

UNITED STATES OF AMERICA  
DEPARTMENT OF HEALTH AND HUMAN SERVICES  
FOOD AND DRUG ADMINISTRATION

+ + +

CENTER FOR DEVICES AND RADIOLOGICAL HEALTH

+ + +

FDA/AMERICAN GLAUCOMA SOCIETY WORKSHOP  
ON THE VALIDITY, RELIABILITY, AND USABILITY  
OF GLAUCOMA IMAGING DEVICES

+ + +

October 5, 2012  
8:00 a.m.

FDA White Oak Campus  
10903 New Hampshire Avenue  
Building 31 Conference Center  
Great Room, Sections B & C  
Silver Spring, MD 20993

MODERATORS: JEFFREY M. LIEBMANN, M.D.  
President, American Glaucoma Society (AGS)

MALVINA B. EYDELMAN, M.D.  
Director, Division of Ophthalmic, Neurological and Ear,  
Nose and Throat Devices, Office of Device Evaluation  
CDRH/FDA

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

WELCOME AND INTRODUCTIONS:

JEFFREY M. LIEBMANN, M.D.  
President, American Glaucoma Society (AGS)

MALVINA B. EYDELMAN, M.D.  
Director, Division of Ophthalmic, Neurological and Ear, Nose and Throat  
Devices, Office of Device Evaluation  
CDRH/FDA

WILLIAM H. MAISEL, M.D., M.P.H.  
Deputy Director for Science  
CDRH/FDA

ROHIT VARMA, M.D., M.P.H.  
Professor of Ophthalmology  
Director, Glaucoma Service, Ocular Epidemiology Center, and the Clinical  
Trials, Keck Medical Center  
University of Southern California

JOSHUA D. STEIN, M.D., M.S.  
Assistant Professor, Ophthalmology and Visual Sciences  
University of Michigan Kellogg Eye Center

DAVID S. FRIEDMAN, M.D., M.P.H., Ph.D.  
Professor, Dana Center for Preventive Ophthalmology  
Wilmer Eye Institute at Johns Hopkins

GLAUCOMA STRUCTURAL DIAGNOSTIC TECHNOLOGIES:

KULDEV SINGH, M.D., M.P.H.  
Professor, Med Center Line, Ophthalmology  
Byers Eye Institute at Stanford  
Stanford School of Medicine

HENRY D. JAMPEL, M.D., M.H.S.  
Odd Fellows and Rebekahs Professor of Ophthalmology  
Glaucoma Division  
Wilmer Eye Institute at Johns Hopkins

GEORGE A. CIOFFI, M.D.  
Chair, Department of Ophthalmology  
Director, Edward S. Harkness Eye Institute  
Columbia University

MURRAY FINGERET, O.D.  
The Glaucoma Foundation

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

JOEL S. SCHUMAN, M.D.  
 Eye & Ear Foundation Professor and Chairman, Department of Ophthalmology  
 Director of UPMC Eye Center  
 Professor of Bioengineering  
 Glaucoma Service  
 University of Pittsburgh School of Medicine

TERESA C. CHEN, M.D.  
 Associate Professor of Ophthalmology  
 Harvard Medical School  
 Chief Quality Officer of Ophthalmology, Glaucoma Service  
 Massachusetts Eye and Ear Infirmary

CHARACTERIZATION OF OCT AS A MEASUREMENT TOOL:

DAVID S. GREENFIELD, M.D.  
 Bascom Palmer Eye Institute  
 University of Miami Health System

GADI WOLLSTEIN, M.D.  
 Associate Professor of Ophthalmology  
 Director, Ophthalmic Imaging Research Laboratories  
 University of Pittsburgh School of Medicine

GIANMARCO VIZZERI, M.D.  
 Associate Professor of Ophthalmology  
 Director, Ophthalmology Clinical Research Center  
 Director of Quality and Safety  
 The University of Texas Medical Branch at Galveston

DONALD C. HOOD, B.A., M.Sc., Ph.D.  
 James F. Bender Professor  
 Columbia University

GENE HILMANTEL, O.D., M.S.  
 Division of Ophthalmic, Neurological and Ear, Nose and Throat Devices  
 Office of Device Evaluation  
 CDRH/FDA

CHRISTOPHER A. GIRKIN, M.D., MSPH  
 Chairman, Department of Ophthalmology  
 University of Alabama at Birmingham, School of Medicine

DONALD L. BUDENZ, M.D., M.P.H.  
 Professor and Chairman, Department of Ophthalmology  
 The University of North Carolina at Chapel Hill School of Medicine

Free State Reporting, Inc.  
 1378 Cape Saint Claire Road  
 Annapolis, M.D. 21409  
 (410) 974-0947

DOUGLAS J. RHEE, M.D.  
 Associate Chief, Operations & Practice Development and Medical Director,  
 Stoneham & E Bridgewater  
 Massachusetts Eye and Ear Infirmary

OCT NORMATIVE DATABASES: CONSTRUCTION, RELIABILITY, AND USABILITY

MALIK Y. KAHOOK, M.D.  
 Professor of Ophthalmology  
 University of Colorado

SHAN C. LIN, M.D.  
 Associate Professor of Clinical Ophthalmology, Department of Ophthalmology  
 University of California, San Francisco

ROBERT D. FECHTNER, M.D.  
 Professor, Institute of Ophthalmology and Visual Science  
 New Jersey Medical School

KOUROS NOURI-MAHDAVI, M.D., M.Sc.  
 Assistant Professor of Ophthalmology  
 Jules Stein Eye Institute  
 University of California, Los Angeles

LINDA M. ZANGWILL, Ph.D.  
 Diagnostic Imaging Laboratories  
 Hamilton Glaucoma Center  
 Associate Professor, Department of Ophthalmology  
 University of California, San Diego

RICHARD K. LEE, M.D., Ph.D.  
 Associate Professor of Ophthalmology  
 Bascom Palmer Eye Institute  
 University of Miami Health System

GUSTAVO V. DE MORAES, M.D.  
 Research Associate Professor  
 Department of Ophthalmology  
 NYU Langone Medical Center

KRISTEN L. MEIER, Ph.D.  
 Division of Biostatistics  
 Office of Surveillance and Biometrics  
 CDRH/FDA

Free State Reporting, Inc.  
 1378 Cape Saint Claire Road  
 Annapolis, M.D. 21409  
 (410) 974-0947

OCT DIAGNOSTIC PERFORMANCE FOR GLAUCOMA

ROBERT N. WEINREB, M.D.  
Chairman, Department of Ophthalmology  
Director, Hamilton Glaucoma Center  
University of California, San Diego

DALE K. HEUER, M.D.  
Professor and Chairman, Department of Ophthalmology  
Director, Eye Institute  
Medical College of Wisconsin

FELIPE A. MEDEIROS, M.D., Ph.D.  
Professor of Ophthalmology  
Hamilton Glaucoma Center  
University of California, San Diego

ROBERT T. CHANG, M.D.  
Assistant Professor - Med Center Line, Ophthalmology  
Byers Eye Institute at Stanford  
Stanford School of Medicine

SANJAY G. ASRANI, M.D.  
Division of Glaucoma Ophthalmology  
Department of Ophthalmology  
Duke University

ANNE L. COLEMAN, M.D., Ph.D.  
Fran and Ray Stark Foundation Professor of Ophthalmology  
Director, Center for Eye Epidemiology  
Jules Stein Eye Institute  
University of California, Los Angeles

EVA RORER, M.D.  
Division of Ophthalmic, Neurological and Ear, Nose and Throat Devices  
Office of Device Evaluation  
CDRH/FDA

HENRY D. JAMPEL, M.D., M.H.S.  
Odd Fellows and Rebekahs Professor of Ophthalmology  
Glaucoma Division  
Wilmer Eye Institute at Johns Hopkins  
OTHER APPEARANCES:

DOUG ANDERSON, M.D.  
Consultant  
Carl Zeiss Meditec

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

ERIC L. BUCKLAND, B.S., M.S., Ph.D.  
Director, President, and Chief Executive Officer  
Bioptigen

PEPE DAVIS, Ph.D., M.S.  
Vice President, Process Excellence  
Carl Zeiss Meditec

MIKE SAND  
Topcon

STEPHEN H. SINCLAIR, M.D.  
Biometrics  
Intelligent Medical Imaging Technologies

MICHAEL PATELLA  
Vice President, Professional Affairs  
Carl Zeiss Meditec

JIN JIL  
Optovue

CARLO TRAVERSO, M.D.  
Professor, Department of Neurosciences, Ophthalmology and Genetics  
University of Genova

LCDR BRAD CUNNINGHAM  
Branch Chief, Ophthalmic Lasers, Neuromuscular Stimulators and  
Diagnostic Devices  
CDRH/FDA

## INDEX

## PAGE

## WELCOME AND INTRODUCTIONS

Moderators: Jeffrey M. Liebmann, M.D. and Malvina B. Eydelman, M.D.

Opening Remarks/Welcome to FDA - William H. Maisel, M.D., M.P.H.	10
Welcome and Introduction Malvina B. Eydelman, M.D.	14
Jeffrey M. Liebmann, M.D.	16
Pearls and Pitfalls of Early Glaucoma Detection The Benefit of Early Detection - Rohit Varma, M.D., M.P.H.	19
Public Health Concerns (false positives and false negatives) - Joshua D. Stein, M.D., M.S.	23
The Impact of Current Diagnostic Practice Patterns (clinical reference methods) on the Evaluation of New Diagnostic Technologies - David S. Friedman, M.D., M.P.H., Ph.D.	30
Current Regulatory Pathways for Glaucoma Imaging Devices - Malvina B. Eydelman, M.D.	36

## GLAUCOMA STRUCTURAL DIAGNOSTIC TECHNOLOGIES

Moderators: Kuldev Singh, M.D., M.P.H. and Henry D. Jampel, M.D., M.H.S.

Detection of Glaucomatous Structural Damage Qualitative Optic Disc Photography - George A. Cioffi, M.D.	41
Quantitative: HRT - George A. Cioffi, M.D.	45
GDx - Murray Fingeret, O.D.	48
Introduction to Optical Coherence Tomography (OCT) Technology Time-Domain and Spectral-Domain Including Similarities and Differences between these Technologies - Joel S. Schuman, M.D.	52
Clinical Applications of SD-OCT (diagnosis, risk stratification, longitudinal assessment) - Teresa C. Chen, M.D.	57

## INDEX

## PAGE

## CHARACTERIZATION OF OCT AS A MEASUREMENT TOOL

Moderators: David S. Greenfield, M.D. and Gadi Wollstein, M.D.

Factors that Impact Image Quality and Measurements  
 How to Assess Reliability and Quality of a Scan, Patient  
 Dependent Factors: Effect of Pupil Size, Media Opacity,  
 and Anatomic Variations (e.g., optic nerve tilt, peripapillary  
 atrophy) on Measurements, Operator Dependent Factors:  
 Centration, Alignment (e.g., head tilt), Instructions -  
 Gianmarco Vizzeri, M.D. 64

Instrument dependent factors (Image segmentation,  
 image registration) - Donald C. Hood, B.A., M.Sc., Ph.D. 71

Measurement Validation: Considerations in Premarket Review -  
 Gene Hilmantel, O.D., M.S. 79

Measurement Validation: Terminology and Concepts -  
 Gene Hilmantel, O.D., M.S. 85

Validation (Agreement and Precision) of Peripapillary Nerve  
 Fiber Layer, Optic Nerve Head and Macula-Based Measurements  
 (GCC, RNFL and total macular thickness) -  
 Christopher A. Girkin, M.D., MSPH 87

Variability of Peripapillary Nerve Fiber Layer and Optic Nerve  
 Head Measurements - Implications for Detecting Change -  
 Donald L. Budenz, M.D., M.P.H. 100

PANEL AND OPEN MICROPHONE DISCUSSION 110

OCT NORMATIVE DATABASES: CONSTRUCTION, RELIABILITY, AND  
USABILITY

Moderators: Malik Y. Kahook, M.D. and Shan C. Lin, M.D.

Review of Normative Database Construction in Available OCT  
 Models Highlighting Differences - Robert D. Fechtner, M.D. 129

Statistical Approaches for Determining Normal Limits in  
 Database - Kouros Nouri-Mahdavi, M.D. 139

Stratification of Normative Data - Linda M. Zangwill, Ph.D. 146

Review of how Normative Database Information is Displayed  
 by Software/Printouts and Described in Labeling for Various  
 Models - Richard K. Lee, M.D., Ph.D. 153

Free State Reporting, Inc.  
 1378 Cape Saint Claire Road  
 Annapolis, M.D. 21409  
 (410) 974-0947

## INDEX

	PAGE
How does Construction of and Statistical Modeling within OCT Normative Databases Compare with Standard Automated Perimetry Databases? - Gustavo V. de Moraes, M.D.	162
Normative Database Construction, Reliability, and Usability: Considerations in Premarket Review - Kristen Meier, Ph.D.	170
PANEL AND OPEN MICROPHONE DISCUSSION	179
OCT DIAGNOSTIC PERFORMANCE FOR GLAUCOMA Moderators: Robert N. Weinreb, M.D. and Dale K. Heuer, M.D.	
Considerations for Evaluation of Diagnostic Performance - Felipe A. Medeiros, M.D., Ph.D.	198
Diagnostic Performance Studies	
Peripapillary Nerve Fiber Layer - Felipe A. Medeiros, M.D., Ph.D.	210
Optic Nerve Head Measurements - Robert T. Chang, M.D.	217
Macular Measurements - Sanjay G. Asrani, M.D.	224
“Screening”: Can OCT Imaging be Used to Differentiate Normal from Glaucomatous Eyes in the General Population? - Anne L. Coleman, M.D., Ph.D.	231
Considerations in Characterizing the Diagnostic Performance of a Device - Eva Rorer, M.D.	237
PANEL AND OPEN MICROPHONE DISCUSSION	245
CLOSING COMMENTS FROM CO-SPONSORS:	
Summary of Workshop Discussion and Next Steps for AGS	260
Summary of Workshop Discussion and Next Steps for FDA	260
ADJOURNMENT	261

M E E T I N G

(8:00 a.m.)

DR. LIEBMANN: Well, good morning, everyone, and welcome to our FDA/AGS workshop on glaucoma databases for imaging devices.

Our first speaker this morning, to kick us off, is Dr. William Maisel, who's a Deputy Director of the Center for Devices.

Dr. Maisel.

DR. MAISEL: Good morning. Nice to see so many faces smiling. And I've learned to judge the audience by how close to the front or how far to the back they sit. So it looks like I can already tell we have a very engaged group this morning.

Welcome to the FDA and to this workshop on the validity, reliability, and usability of glaucoma imaging devices. And the workshop we're co-sponsoring, as you know, with the American Glaucoma Society.

It really is great to see such a big turnout, and we have additional people on the web watching. And I'm sure the original 13 people who founded the American Glaucoma Society in 1985 would be really thrilled to see how many people are here today and interested in the topic.

It's obviously, as you know, an enormously important topic, with glaucoma the second leading cause of blindness not only in the United States but worldwide. And at FDA, the Center for Devices and Radiological Health, or CDRH, is responsible for assessing the safety and effectiveness of

medical devices, including both therapeutic and diagnostic devices, as well as radiation-emitting products. And so, certainly, this is a topic of great interest to us.

Our mission includes not only a responsibility to protect the public health but also to promote the public health, and that latter aspect is something we've really been focused on over the last two or three years. We've launched a number of efforts to help bring high-quality, safe, and effective devices to patients, and we're very focused on getting those devices to market in a timely manner.

We recognize we can't do this alone, and in many respects, being here today is emblematic of that. We really feel we need to understand the science, engage with the clinical community, industry, the patient community, so that we can find the sweet spot, if you will, and really advance the field. It's critical that we hear from people outside the Agency. And we have other efforts, not just workshops like this, but other efforts to engage actively with people outside the Agency, and I'll highlight just a couple this morning.

Last year we launched our Entrepreneurs-in-Residence program, which, in collaboration with the White House Office of Science and Technology Policy, brought in outside experts into FDA, into CDRH. And we sat side by side with people from industry, other government agencies, academia, and took a hard look at how we do business, how we evaluate

products, looking for increased efficiencies and thinking of ways that we could maintain our scientific rigor, but do things more quickly and help promote the development of products. And as an outcome of that process, earlier this year we announced our Innovation Pathway 2.0, which currently is focused on end-stage renal disease products of a certain type.

But really the lessons we learn from evaluating these products and from applying these new ideas and new tools will expand more broadly as we get comfortable with them and as we find the ones that work and discard the ones that don't.

We've also developed what we've called our Network of Experts, which is really a place where our staff, our scientific and review staff, can go to people outside the Agency, to professional societies, other experienced people and experts, to get input on scientific ideas, to better understand the products and technologies that are being developed, so that we can quickly get input from outside the Agency. We have an incredible staff that are very dedicated, knowledgeable, and experienced, but that's our staff, and we recognize that there are other viewpoints and other very knowledgeable people. And we're really trying hard to incorporate those ideas into our policies and practices.

So on many levels, today's workshop is really emblematic of a broader effort going on in many device areas, in many disease areas, throughout the Center.

Today's workshop obviously is focusing on novel imaging devices, and they've really transformed how many diseases, not just glaucoma, are diagnosed and in many cases treated. We've seen an exponential rise in different types of innovative diagnostic technologies, and one of them that will be discussed today, obviously, is optical coherence tomography.

It truly takes a village, and not only including stakeholders outside the Agency, but we have a number of knowledgeable people in different organizational units within our own Center that collaborate not only to put on a workshop like this, but to evaluate these products.

So we have our Office of Device Evaluation, which is the premarket office that evaluates these products. We collaborate with our Office of Surveillance and Biometrics that has statistical expertise. We have an Office of Science and Engineering Laboratories that develops test methodologies and studies these products and devices so that we can make sure we're asking the right questions, that we don't ask questions that aren't important, and that we get the information we need in the least burdensome way possible. So it really does take a big collaboration.

And the input we obtain today will certainly go a long way towards maintaining the scientific rigor in evaluating these products, but also getting the right products to the right patients and putting tools in your hands that you can use to take better care of your patients.

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

So on behalf of FDA and CDRH, I want to thank you very much for attending, and we look forward to a productive workshop.

(Applause.)

DR. LIEBMANN: Thank you, Dr. Maisel, who the AGS is pleased to work with for today's meeting, but also in the Network of Experts process as that evolves over the next few months.

Our next speaker is well known to the AGS, members in the audience. She even speaks at our annual meeting periodically. Dr. Malvina Eydelman is the Division Director for Ophthalmic, Neurological, and ENT Devices.

Malvina.

DR. EYDELMAN: Thank you, Jeff.

Welcome, everybody. We're delighted that all of you have made it to White Oak, made it through security and are actually here joining us.

Well, everybody in this audience is well aware of the fact that open-angle glaucoma affects an estimated 2.2 million people in the U.S., and the number is likely to increase to 3.4 million in 2020, with 130,000 in the U.S. blind from glaucoma. All of that underscores the impact of glaucoma on public health.

The last decade has seen an incredible increase in the use of imaging devices in glaucoma patients. OCT is leading the way with its

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

exponential utilization.

With these devices being increasingly used in clinical practice for clinical decision making and management, we at FDA are seeing glaucoma imaging devices becoming the focus of significant premarket activity. Additionally, imaging device parameters are increasingly being used as part of the enrollment criteria in glaucoma therapeutic studies.

All of us, as a society, are becoming overly dependent on technology. Likewise, the ophthalmic community has grown to be overly dependent on the use of imaging devices in glaucoma diagnosis and treatment, and to highlight this, I would like to read a couple of quotes from the recent literature articles.

When clinicians observe red lettering in data printouts or see red numbers out of the normal range, the presumption is that those results are abnormal. We have become enamored with sophisticated analysis algorithms and colorful printouts before we have studies that show what the results of the tests mean. Although there is objective image, subjective quality assessment is still necessary to ensure that the image acquired is adequate for evaluation and that the analysis algorithm has functioned properly.

In light of this, we hope that today's workshop will highlight the issues that the user should be aware of when interpreting information from glaucoma diagnostic devices.

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

Today's workshop will provide industry insight into current FDA thinking. And last, but not least, we're hoping to receive valuable input from academia, industry, and other stakeholders on how to improve our regulatory science, specifically how to fine-tune endpoints and strategies for assessing the relative safety and effectiveness of this new diagnostic information in the management of glaucoma.

And I'm certain that today's workshop will be a great start to the challenge that was recently posed by Stein et al. in their *Ophthalmology* 2011 article, that great attention needs to be paid to more fully understand and overcome some of the known limitations associated with current ocular imaging devices.

Thank you.

(Applause.)

DR. LIEBMANN: Well, on behalf of the American Glaucoma Society, I'd also like to extend our welcome to everyone who's here today.

These are my disclosures in the handouts.

The mission of the American Glaucoma Society is to promote excellence in the care of patients with glaucoma and preserve or enhance vision by supporting glaucoma specialists and scientists through the advancement of education and research. We view this as a collaboration between glaucoma clinicians, clinician-scientists, our industry colleagues, and of course our regulatory colleagues. The AGS membership is rapidly

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

expanding, as are the number, of course, of glaucoma patients. We have now more than 900 members.

Glaucoma, as you'll hear today, is an optic neuropathy characterized by loss of retinal ganglion cells and their axons. That creates both a structural and functional deficit related to the optic nerve, and we'll talk about the structural component today.

Imaging has become critical, really, to the evaluation of the glaucoma patient and provides quantitative information about optic nerve structure. In clinical practice, we can use imaging to stage disease, perhaps better assess glaucoma suspects for the presence of disease, aid in the detection of glaucoma progression, and in many studies, bring the typical clinician to the level of the expert observer with respect to diagnosis and progression.

In research, we use it to study longitudinal structure function changes of the optic nerve. We use it to determine endpoints and perhaps use it as a way to determine which patient should be included in clinical trials. Certainly these devices are beginning to enjoy widespread use, not only in the United States, but around the world.

The purpose of this meeting, as listed in your program, is to provide a platform for us to discuss with our FDA colleagues all the issues surrounding these imaging devices and their databases. We certainly want to make certain that these databases are used appropriately in clinical practice,

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

and they're evaluated over time to determine how well they should be used in practice to enhance the health of our patients. And lastly, as Malvina mentioned, we all want to help advance the regulatory science surrounding these devices.

So please participate in the discussion and the panels. This is your opportunity to provide feedback. And eventually we'll create a manuscript from this meeting, to memorialize it and move forward to other types of meetings.

I want to thank all of the speakers who took time from their busy schedules to come down to Silver Spring today to help us, and particularly the members of the organizing committee, who are listed here on the screen and will help with moderating these sessions.

Lastly, I want to thank the AGS itself and also the American Glaucoma Society Foundation for helping to underwrite the cost of this meeting.

Thank you very much.

(Applause.)

DR. LIEBMANN: One last hitch. We'll continue this conversation at our annual meeting in San Francisco.

Thank you.

DR. EYDELMAN: Thank you, Jeff.

Now, Dr. Rohit Varma will highlight for us "The Benefit of Early

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

Detection."

DR. VARMA: I just wanted to thank the organizers for inviting me.

My task is essentially to outline what's the benefit of early detection and treatment of individuals with this blinding disease.

At the outset, in fact, I just want to state that I don't have any financial disclosures related to this talk.

So the goal, in essence, of detection and treatment is to try and prevent people from getting functional loss and to have a negative impact on their ability to do everyday tasks. That's the ultimate goal of why we treat people, why we detect early damage in this disease. And so it's important to assess that early change in an individual's ability to do everyday tasks.

The way that one assesses it is by using what are called patient-reported outcome measures, because these are sort of formal ways of assessing how an individual is doing their everyday tasks and whether they feel that they are able to do those tasks or not. So one gets an idea from the patient's standpoint as to how they are performing. In addition to various tests which you'll hear about today, it's critical to evaluate how an individual is doing, and this is how one assesses it.

There are various instruments which have been used. One of them, which has a fairly good reliability, both internal and external, is the NEI-VFQ, which has been used in a whole series of assessments, both in large

groups of individuals in trials and in small case series. And the advantage of this instrument is that it's able to detect change over time, as well as, in fact, one can interpret what's a minimally important change.

So one of the issues which comes to the mind is what minimum amount of change is meaningful, in terms of visual field loss, to an individual's everyday ability to do their tasks, and what kinds of activities are impacted.

One of the assessments that we did was on a large group of individuals in LA who, very importantly, did not have any idea that they had disease, and we assessed them both in terms of their visual field and in terms of their ability to do everyday tasks. And one of the things which we found was that, even with early visual field loss -- and that's to the left of this straight line, where you see a mean dB of two to four loss, in fact, in the better-seeing eye -- you begin to see an impact on an individual's ability to do tasks. They begin to have problems, they begin to have difficulty doing everyday tasks, particularly driving, in this instance, which is the dashed line you see, which has this almost linear relationship as it goes down.

Now, this is with the better-seeing eye. But if you look at data where both eyes are involved, where visual field loss in both eyes are assessed, you see the same general pattern, that people who have minimal loss or early loss, in fact, see an impact on their ability to do everyday tasks and begin to have problems. That pattern persists for a whole series of tasks.

Now, this graph shows you, not sort of the mean change with

visual field loss, but the proportion of individuals that begin to have difficulty. And on the X-axis, to the left, you see individuals with minimal loss. To the right are individuals with more loss. And as you can see, in fact, that even with minimal loss, you begin to see that people increasingly have difficulty doing, for example, in this graph, driving at night, at reading ordinary print, at reading street signs. Pretty much across the board, in fact, they have a whole series of activities which are involved with their everyday lives that would begin to have an impact with minimum field loss. And here, as well, is dependence on others.

Now, in addition to that, there have been other studies which have assessed ability to walk outside or do other tasks within the house, and you begin to see an impact again when people have mild visual field loss.

Now, those assessments which you saw are on a cross-sectional level. But when you look at change in visual field over time, the same pattern persists, which is that, to the right over here, you see greater field loss, to the left there's less, and as one begins to have more and more change in visual fields, you begin to see a greater impact on an individual's ability to do those tasks. So loss in ability to do tasks is seen even with those with mild visual field loss and a three to four dB, not just difference, but even progressive change is associated with impact on various activities, driving being the most sensitive of them.

But when one looks overall at field loss, yes, it's true. But what

predicts getting greater field loss? And one of the trials which has been done in Sweden, which looked at individuals with early damage, showed that having minimal field loss at the outset predicted going on to get greater field loss over time. And so it's important to try and prevent early visual field loss, because those people that get early field loss go on to have progressive damage over time, even when they're treated.

But it's not just true about early visual field loss. What's also important -- and from the OHTS study, which many of you all know are people that did not have damage but were at high risk of developing optic nerve damage and visual field loss. What the OHTS study showed was that structural changes, even structural changes early on, predict which individuals go on to get visual field loss and optic nerve change over time.

And so, clearly, in fact, it's important not just to identify individuals who have early field loss but actually go a step earlier, which is identify individuals who have structural loss, because those then predict which individuals will go on to have field loss and ultimately have an impact on their everyday lives.

So clearly, in fact, from the data which exists from large studies, from trials and so on, it's clearly important to try and prevent the development of early damage to the optic nerve. And I'm sure you'll hear more about it as the day goes on.

So I appreciate the invitation, and I thank you very much.

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

DR. EYDELMAN: Thank you very much.

(Applause.)

DR. EYDELMAN: I would like to ask Dr. Joshua Stein to summarize for us "Public Health Concerns."

DR. STEIN: Good morning. Thank you for inviting me to speak at this workshop.

Diagnosing patients with early glaucoma can be challenging. It's estimated, of the two to three million Americans with glaucoma, approximately half of them do not even know they have the disease. Complicating matters further are findings out of Australia, from Hugh Taylor's group, demonstrating that 50% of patients who are newly enrolled in a large population-based study had actually -- that were newly diagnosed with glaucoma upon study entry had actually seen an eye care provider the year before the study and the provider missed the diagnosis.

So why is it challenging to identify patients with early glaucoma? It was once thought that patients could be identified with glaucoma based on elevated intraocular pressure, a known risk factor for glaucoma. But Harry Quigley and others have found that 20% to 50% of glaucoma patients actually have pressures that are in the range of normal most of the time.

Visual field testing is usually thought of as the gold standard at identifying patients with glaucoma. Yet, as Dr. Varma pointed out, a lot of

structural damage can occur before changes occur on the visual field. In addition, this test is not easy for a number of patients to take.

Researchers have found considerable interobserver variability in cup-to-disc estimates among groups of optometrists, ophthalmology residents, and even glaucoma specialists. Thus, there's a clear need for alternative ways to help clinicians diagnose glaucoma and monitor for disease progression.

Benefits of ocular imaging devices include an ability to identify structural damage to the nerve and nerve fiber layer sooner than can be detectable on perimetry. These devices are patient friendly and, unlike with perimetric testing, do not require subjective patient input. They often can be performed without requiring dilation, and the color coding system for abnormalities can make them relatively easy for providers to interpret.

My colleagues and I recently studied the utilization of ocular imaging devices among enrollees with glaucoma in a large matched network throughout the United States and found a dramatic rise in the use of these tests over time.

The integration of ocular imaging devices into practice has led to several important public health considerations, such as whether these devices have adequate sensitivity and specificity, how valuable these tests are at identifying patients who have glaucoma, and whether minorities are getting access to these technological innovations.

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

Prevent Blindness America, an organization focused on ocular health and safety, recommends, for an instrument to be adopted for screening for glaucoma, it should meet minimum standards of 85% sensitivity and 95% specificity for detecting moderate to severe glaucoma. They didn't recommend any cutoffs for earlier pre-perimetric glaucoma.

When evaluating estimates of sensitivity and specificity, it's important to keep in mind that if standard automated perimetry is being used as the gold standard in determining the presence or absence of disease, some patients with early glaucoma may still have normal perimetry, and thus, that may impact the sensitivity and specificity estimates.

Comparing the various studies in the literature that report sensitivity and specificity of OCT and other ocular imaging devices can be challenging because studies vary with respect to the specific device and model tested, the parameters studied, and the characteristics of the study participants.

Keeping those considerations in mind, here are some selected study results.

Using the less stringent 5% cutoff to identify glaucomatous damage, studies report sensitivities in the mid to high 80s for early to moderate glaucoma, and up into the high 90s for advanced glaucoma. With the more 1% stringent threshold, as one might expect, the sensitivity of these devices go down while the specificity goes up.

Using computer algorithms to identify the best set of parameters to distinguish normal from glaucomatous damage, researchers have been able to achieve sensitivity and specificity levels in the mid to high 90s.

Comparisons between the Cirrus and Stratus OCT units have found, although there's excellent correlation between the nerve fiber layer thicknesses between these devices, they actually differed significantly in their normative classification among the eyes studied.

Another study concluded that direct comparisons between nerve fiber layer measurements between the two units were not comparable. These findings can have important implications for clinical practice if patients transfer care from one provider, who's using one device, to another provider, who's using a different device.

Occasionally glaucoma specialists will encounter patients who are sent in to see us with a condition that some refer to as red disease. These are patients for whom they have a completely healthy optic nerve and normal visual field, yet the imaging shows a lot of areas of red or abnormalities.

Here's a patient of mine who fits this profile.

From a public health perspective, mischaracterizing patients as having glaucoma when they actually do not have disease can be problematic because it can result in patient fear and anxiety that they have a condition that can lead to irreversible blindness, the subjection of patients to

unnecessary interventions and potential side effects, and unnecessary costs.

Likewise, mischaracterizing patients as disease free when they indeed have glaucoma is also problematic. If a patient is mistakenly classified as normal, they may leave the clinic thinking they have a false sense of security and are less likely to follow up for return care until more damage has occurred.

Another public health concern pertains to the information gained from performing ocular imaging. If the pre-test and post-test probability of a patient being characterized as having glaucoma is the same, then the test itself is yielding very little useful information.

Medeiros and colleagues found that abnormal findings on OCT led to a large change in the likelihood of a patient having glaucoma, while it was less useful for those who were characterized as normal by the test.

With the rapid increase in utilization of ocular imaging in glaucoma management, another issue that needs to be considered is the extent by which eye care providers are using imaging devices as an adjunct to perimetry and examination of the nerve, or whether they're using imaging instead as a replacement for some of the more traditional tests.

In a study conducted by my colleagues and I, we learned that not only was there a rise in the use of ocular imaging, but there was also a pretty dramatic decline in the probability of a patient with glaucoma undergoing visual field testing.

These are four patients being seen exclusively by ophthalmologists.

And here are patients under the care of optometrists exclusively, and they actually had a higher probability of undergoing imaging and visual field testing.

Here's a patient that I saw in clinic last week. Their referring doctor was concerned about progressive worsening of glaucoma on OCT and requested a formal glaucoma evaluation. When I requested copies of prior visual fields, their office told me the last field they had done was two years earlier. When I performed the visual field test on the patient, I found what appears to be an evolving bitemporal hemianopic defect concerning for a lesion of his chiasm. This case highlights the types of problems that can arise when providers use imaging as a replacement for perimetry and careful examination of the optic nerve.

Finally, we need to be sure that these innovative technologies are being made available to patients of different sociodemographic profiles. Here we assess whether racial disparities exist in the receipt of ocular imaging devices. And despite the fact that all of the patients in the study had health insurance, we identified that Latinos had a much lower probability of undergoing imaging, relative to some of the other racial groups.

It's my hope, in the future, that we'll see improvements to the normative databases for ocular imaging devices so there are larger numbers

and improved diversity of patients included. This will certainly enhance their ability for the devices to identify those with and without disease.

I also believe it will be very important for device companies to fully disclose the number and characteristics of the types of patients in their normative databases.

Third, there needs to be improved compatibility between different OCT devices and models, so patients can be followed longitudinally to monitor for glaucoma progression even if the provider upgrades his equipment or the patient changes to another provider using a different unit.

Fourth, it is essential that the output generated from ocular imaging devices are open source so providers can extract individual data elements and integrate them into electronic health records and so researchers can obtain access to this sort of data.

Fifth, we need to be sure that patients have access to technological innovations, irrespective of their race and socioeconomic status, and finally, that providers need to be further educated on the strengths and limitations of ocular imaging devices and how they can be an excellent adjunct to, but not a replacement for, some of the more traditional tests.

So, in conclusion, from a public health standpoint, there's a substantial benefit from ocular imaging devices in identifying patients with glaucoma. Hopefully this conference will highlight ways to further advance

the science so we can overcome some of the existing limitations of these devices.

Thank you.

(Applause.)

DR. LIEBMANN: Thank you, Josh, for highlighting some of the challenges that all diagnostic devices face.

As we continue this series of introductory talks, David Friedman from Hopkins is going to try to address the question of how we deal with an emerging technology that might be better, perhaps, as a reference standard than the current SAP reference standard.

David.

DR. FRIEDMAN: The current reference standards used to monitor and treat patients with glaucoma are largely assessing visual field and visual field worsening. And that's in large part due to the fact that we know the impact of visual field loss on function, and therefore it's a very good surrogate for disability and further problems for the patient.

Other important reference standards include optic nerve head changes. And that's been used in several clinical trials, which I'll briefly highlight in a moment. And recently we've had, as people have been referencing, development of newer technologies that also hold promise to be used as reference standards.

This is a just a simple graphic from a summary review by

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

Dr. Malik, showing that in pre-glaucoma in the Ocular Hypertension Treatment Study, a large proportion of individuals progressed on the basis of nerve findings prior to there being any detectable change in the visual field. All of these patients had normal visual fields. But in this case, because it was prior to the development of glaucomatous field loss, the nerve head could be used as a way to detect change. This is a problematic reference standard. I'm going to talk about that in a minute. But in this kind of situation where you don't yet have field loss, the nerve head can work.

In glaucoma patients in the Early Manifest Glaucoma Trial, it was a very different methodology, and there may be reasons why so few were identified as having nerve head changes. But, largely, you see the visual field as being the reference standard in this kind of situation that identifies worsening of visual field.

The visual field is clearly a very important endpoint. It documents factors that are associated with worse function for patients, and moderate bilateral visual field loss is clearly associated with declines.

Rohit pointed out some data from the Latino study in Los Angeles. This is from African Americans and whites, older, who were in a population they studied on the Eastern Shore of Maryland, showing quite similar findings using the ABVF, not the NEI-VFQ. We have the slide that Rohit just showed a moment ago.

But there are problems with these visual function questions.

This is from the Collaborative Initial Glaucoma Treatment Study by Jampel and colleagues, and what it shows is a fairly low correlation between self-report and not only visual field but also visual acuity. And what is really striking in this study is a patient's sense of their mood or burden to others actually dominates the visual issues in visual complaints. And so I am worried a little bit about an over-reliance on patient-reported outcomes in these kinds of populations who have a high prevalence of these depressive symptoms.

And instead we focused recently, and others as well, on the impact of field loss on function. Here's one example where bilateral field loss is associated with much slower sustained reading, as you can see, comparing the red controls to the blue cases, with an average decrease of seven words per minute per five decibels of field loss. And it was linear.

If you look at steps per day, we measure people wearing an Actical monitor that monitors actual steps. You can see, as you have increasing field loss, you have fewer steps per day. People with field damage do less.

And here is a prospective study showing that those who have field loss are more likely to stop driving, to lose their independence, which ultimately has enormous ramifications for their ability to live freely in the community.

But visual fields, while they're very good at documenting loss of

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

function, are not ideal as a reference standard. And there are several reasons for this. There are many cases, and we can show anecdote after anecdote. But there's also research with nice large population studies showing that visual field loss is present at times -- is not present at times when there's clear documented loss of nerve fiber or nerve disease. There can even be worsening of the glaucoma, as documented in other ways, prior to any field loss showing. And it really is difficult, once you have more advanced disease, to rely on the visual field for assessing further worsening of the disease.

This goes way back to Harry Quigley's early work on the nerve fiber layer. This is a picture from one of his patients. This patient clearly has a wedge defect in the inferior area of the nerve in that black and white photo, and you can see that, for two years subsequent to that defect being present, there was no field loss. And here on the third year there is evidence of a field defect. This patient died at this time, and there was an autopsy, and we could see that he was missing 30% of his axons. So a large amount of damage had occurred. It was clearly present in that photo, but no evidence on visual field. This is one of the key limitations of this as a reference standard.

We often see nerve fiber layer predating visual field loss. In a large prospective study looking at patients at risk of glaucoma, 50% of those with visual field loss had nerve fiber layer thinning prior to the field loss, whereas only 8% of those who didn't develop field loss had that thinning.

This is the modern version of that from Chris Leung, who sent

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

me this slide. You can see the ongoing in the second -- the red shows you the ongoing progressive loss of nerve fiber inferiorly, with only borderline visual field changes.

It's not 100%. Here we have a case where there's clearly field change but nothing documentable on the nerve fiber layer. So these tests aren't 100% perfect, and not everybody will develop one or the other first.

One concern I do have, and I want to raise here, is while it's very logical to assume that nerve fiber layer loss results in field loss, we need prospective studies to show that this is a really strong correlation so that we can say yes, this is a good standard that predicts what we know leads to disability, because you could imagine other reasons for nerve fiber layer thinning.

This is the example that I wanted to give of dense field loss being very difficult to follow. This patient, both clinically and in research, it would be extremely challenging to know whether or not this patient gets worse. You could alter the testing strategy, but even then it's frequently very hard to know when patients like this are getting worse.

I want to briefly discuss what I feel is the other currently accepted reference standard, which is optic nerve head change. There are cross-sectional studies that seem to indicate that optic nerve head assessment by an expert is similar in performance at identifying glaucoma to that of most of the devices, although the more recent devices may be slightly

better performing. However, it is very difficult using optic nerve head imaging or photos to detect change late in the disease, and it also difficult, once glaucoma is established, for experts to agree on who is getting worse.

Here's an example of late-stage optic nerve damage. You tell me, all the great clinicians out there, if you know when this person's gotten worse. It would be very hard, even with 30 serial photos over 10 years, to know that this nerve has changed, because there's so little rim tissue remaining.

There was a recent publication where experts graded optic nerves in the archives. Fifty-two percent of the decisions of who had gotten worse, based on nerve photos, required adjudication because there was disagreement.

In our study a few years back, where we had prospective data, we took photos that were taken after and before. We randomized the order. We often predicted the earlier photo as being the later photo. It's a problem trying to grade already existent glaucomatous nerves as getting worse or not.

But optic nerve head analysis does seem to perform somewhat better. And this is one of my last slides. This is from Bal Chauhan. He had a study looking at HRT over a long period of time. Those who developed HRT changes over the first half of that time period, about five years, had three times more likelihood of developing field loss in the subsequent period. The nerve head changes predicted the field changes, and we know that that can

be done with some of these imaging devices.

In conclusion, glaucoma -- I don't know if this has been clearly stated so far -- is a slowly progressive disease. The average patient would take 15 years or more to develop severe field loss at the rates of change that we see. We cannot wait 15 years to understand outcomes for patients in studies and in clinical activity, so we need surrogates. Surrogates are vital to our being able to treat and manage this disease. Using the visual field as the main reference standard has significant limitations, particularly early and late in the disease. And I think the newer technologies really do show tremendous promise in helping us to overcome some of these limitations.

Thank you very much.

(Applause.)

DR. LIEBMANN: Thank you, David.

How does the FDA regulate these devices? Our next speaker, Dr. Eydelman, is going to talk about "Current Regulatory Pathways for Glaucoma Imaging Devices."

Malvina.

DR. EYDELMAN: Thanks, Jeff.

I work for the FDA.

(Laughter.)

DR. EYDELMAN: Before we discuss how we regulate glaucoma imaging devices, I wanted to put this up, a definition of what medical device

is, because it actually changes when you go internationally.

In the U.S., medical device is an entity that diagnoses, cures, mitigates, treats, or prevents a disease or a condition that affects the function or structure of the body and does not achieve intended use through chemical action and is not metabolized.

As you can imagine, there's a huge diversity of medical devices that the Center regulates; therefore, we needed to make some sense of it. And the law gives us the flexibility to calibrate our regulatory approach as to the level of the potential risk posed by the new products.

Hence, all of the devices are classified into three classes, with Class I being the simplest or the lowest risk, Class II, more complex, and Class III being the most complex or the highest risk.

For Class I, general controls are needed. Most Class I devices are now exempt from premarket notification.

For Class II, in addition to general controls, FDA requires special controls, and these may include performance standards, FDA guidance documents, device tracking, a patient registry. Most of Class II devices require premarket notification, or 510(k), to show substantial equivalence to a legally marketed predicate device.

Class III is typically reserved for devices that support or sustain human life, have substantial importance in preventing health impairment or potential unreasonable risk of illness or injury. These kinds of devices require

a PMA, or premarket approval, demonstrating reasonable assurance of safety and effectiveness.

And just to put it in perspective, here's an example of an ophthalmic exam. So you can see, with Class I, something like VA charts and perimeters versus IOLs and excimer lasers being a Class III.

So as I alluded to earlier, the two most commonly used regulatory submissions are 510(k) and PMA, with most Class III going to PMA and not exempt Class I or Class II requiring 510(k).

510(k) is a marketing clearance application which allows us to determine substantial equivalence, once again, to a legally marketed device or a predicate device that we refer to. This kind of application needs to demonstrate substantial equivalence, which means that it has the same intended use and has the same technological characteristics as a predicate, or has the same intended use and has different technological characteristics, and information in the 510(k) submission does not raise new questions of safety and effectiveness and demonstrates it is as safe and effective as the predicate.

A PMA, or a premarket approval, on the other hand, is an application requesting clearance to market, based on sufficient valid scientific evidence providing reasonable assurance that the device is safe and effective for its intended use.

For these kinds of determinations, we'll look at intended

populations, conditions of use, probable benefit to health versus probable reliability, and this needs to be done on valid scientific evidence.

Not to bore you to death, but a short summary of the differences between PMA and 510(k). PMA is an approval versus 510(k) is a clearance. 510(k) is Class II devices and shows that a device is substantially equivalent to a predicate as compared to a legally marketed device versus PMA is reserved for Class III and needs to contain valid scientific evidence to show reasonable assurance of safety and effectiveness.

Well, why am I dwelling on all of this? To get back to what we're here to discuss today, which is diagnostic glaucoma devices, most of them are cleared via the 510(k) process, and some of the examples of the ones that are, are slit lamps, tonometers, fundus cameras, GDx, HRT, ultrasound biometry, and OCT. All of these are cleared via the 510(k) process.

When we look at the 510(k), we make sure that the application contains device description, proposed indications for use, which actually dictates the level of data requirement, performance data, which needs to show demonstration of substantial equivalence to the predicate -- again, the performance data that is required in current submissions is just demonstration of substantial equivalence to the predicate -- and proposed labeling.

I want to make sure that everybody in the audience understands what we're talking about when we say labeling.

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

As far as FDA is concerned, all the user manuals, brochures, websites, and package inserts, et cetera, are considered 510(k) labeling. We want to make sure that these documents describe the diagnostic device performance and appropriate use of the device to mitigate risks. Performance data that's usually presented needs to include all of the considerations for the user, and contraindications, warnings, and precautions are placed in the labeling as well.

For those of you in the audience who might be sponsors and have an idea for a great new diagnostic device and you're not sure how to do it and how to fill out applications, I suggest that you utilize our pre-submission program. It's a program that allows an opportunity for a meeting with FDA, and after a short submission, you can ask us specific questions about preclinical and clinical requirements. And the review goal is 60 days.

Summary. These are the points. I know most people don't pay much attention once I start talking regulatory, but these are the key points that I want to make sure you do remember.

The regulations of glaucoma diagnostic devices is usually via the 510(k) path through a substantial equivalence comparison. We review labeling to ensure that it provides a summary of the performance data to allow the user to understand the limitations of the device and adequately interpret the device output. Our 510(k) summaries, which are available on

our website, describe all the information submitted to make FDA's determination of substantial equivalence.

Clearance of a device as a diagnostic tool should not be misinterpreted to mean that the device can be a standalone diagnostic modality for a specific disease.

Clinicians should avoid misinterpretation of results and possibility of improper diagnosis of a treatment.

Unsupported changes in practice due to over-reliance on diagnostic test results are also very worrisome.

Thank you.

(Applause.)

DR. LIEBMANN: Thank you, Malvina.

I'd like to ask Kuldev Singh and Henry Jampel to join us on the podium to moderate the next session on the devices themselves.

DR. SINGH: Thank you very much, Jeff. Thanks, Jeff.

And I'll introduce the first speaker, who is Jack Cioffi. Jack is actually going to give two talks. He will lead off by speaking on "Qualitative Optic Disc Photography" and then will also tell us about HRT.

Jack.

DR. CIOFFI: Thanks, Kuldev. And thanks for the invitation.

I was asked to set the stage a bit. And it seems so easy that we look at a photograph and we look at the nerve and we know what glaucoma

looks like, and why we're here today is that it isn't that easy, obviously.

Here's a slide that I adapted from the FORGE II presentations.

And if I told you this was a small nerve, and you can see there's some vertical cup elongation, a deep cup, maybe some nasalization of the vessels, the so-called ISNT rule is violated, you look down below and you can see the shaded nerve fiber layer defect. No real peripapillary atrophy, but you do have at least one hemorrhage.

But to get there, to detect glaucoma in this nerve, you did a lot of things in your brain just then, and that is, you used your own internal normative database. You knew something about disc size, and how it affects any cup in a small disc is probably significant. You evaluated the topography, perhaps, if you had this in a stereo photograph. You look for peripapillary atrophy. You did a 360 analysis around the nerve. In your own nerve fiber analysis you look for color and you look for the hemorrhage.

So you did a lot of things, and we're asking machines to do an awful lot of things that we've tried to do in the past, but it's difficult. It's difficult because there's a huge variability of appearance of the size of the nerve and angle of insertion, myopia, comorbidities, et cetera, not to mention that, over time, there's been multiple definitions of what a glaucomatous nerve may look like, as highlighted by the Drance paper below.

So the ideal diagnostic test, then, is something that can separate, even with all of these compounding factors, normals from glaucoma

by some continuous variable or perhaps a set of variables. But unfortunately, in the real world, this is what happens, right? We have overlap of the two populations, and it's that zone of uncertainty that we're trying to separate so that we can say, at least for the diagnosis of the disease, that this is glaucoma and this is normal.

Well, what about the other half of the coin? And this is borrowed by the next speaker, Murray Fingeret, that progression -- and that is, in progression, even in this obvious progression, we had to do a couple things. We had to do all the same things we did with diagnosis, all the same analyses that we did for detection, but then we had some other requirements, right? We had the image register, both an X and Y, but also in Z. We had to align them. We had to make sure that there was no cyclotorsion and then we had to do serial image comparisons over time.

So we're asking machines or ourselves, as diagnosticians, to do a lot of things for both establishing disease and establishing progression.

It's scary to think that almost 20 years ago I said that I believe the promise of optic nerve head imaging, in an editorial, is not really detection, not separating it out, but progression. And we're still struggling with that same issue today.

So where have we gotten and what do we know about photography? Because that's the first topic here. Detecting glaucomatous change with photos is not new science.

There's a paper from 1980, from Peterson et al., looking retrospectively at a set of photographs. And unfortunately, the sensitivity of change over time was very poor.

And this is a paper that David Friedman highlighted earlier. And I pulled this graph or this table from their paper, and that is, when they looked, three experts in glaucoma, Henry, David, and Harry, looked at mixed-up slides over time, several things came out. One is that almost half of the eyes or over a third needed to be adjudicated, needed to be looked at after the fact, and they needed to argue about them.

But probably more concerning is two facts from this paper. One is that the kappa, the overall agreement, was only .2, which is very poor. And you see the conclusion that they made. The other thing is that about 40% of the time they actually mixed up the order. When they didn't know the time sequence of the slides, they said that the earlier slide was actually showing progression.

So where are we with disc photography? I think it's an eroding gold standard, unfortunately. There is reasonable reproducibility, and one thing that we do lose when we aren't looking at the nerves, as Engleman (ph.) and others have pointed out, is we lose true color representation that photographs give us. We can do serial comparisons, but the availability of photography is quickly going away.

Stereo photography acquisition systems are almost

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

nonexistent, the migration to digital systems and storage is getting more cumbersome and more difficult, and there's less and less standardization on how you get stereo pairs, even if you can. And then the final insult here is what I just pointed out from the study from Hopkins, and that is, even in the best of hands, even among experts, stereo comparisons aren't agreed upon.

So I think we're here for a reason today, and the reason is this list right here. We're not doing well with how we've done it for the last several decades.

Let me skip to the other slides, if I can. No, it's the top right. There you go.

And sort of an extension of the first talk is what about how we've been doing with confocal laser scanning? And I was asked to touch on HRT, specifically. Just a refresher. I'm not going to go back through how a confocal scanner works, but I do want to remind you that it does provide, for the first time, quantitative analysis of the topography, sort of the landscape of the surrounding -- the nerve and the cup itself. Unfortunately, it's really just a maximal intensity reflectance image. It's telling us about how light is reflected back and picking the point.

So it gives us a surface map. It has given us some very nice progression. And detection analysis software, which I'll touch on briefly, I think has contributed quite a bit. And we thought, going into this, that it would help us detect glaucoma, and it has, to some extent, helped us detect

progression probably to a lesser extent, although the algorithms have gotten quite sophisticated.

Something that we probably didn't predict, because we didn't have the large multicenter studies initially, was that it would actually contribute to the risk profiling. And I think it's done that quite well, and I'll touch on that.

Why are we doing this? And we'll see this slide in some variation a dozen times today, I'm sure. Why are we here? Because fields alone, function alone, actually mix upwards of 50% or 60% of early diagnoses. And this has been shown. It was highlighted before OHTS. It was highlighted in OHTS by Kass et al.

So what about HRT? The progression analysis software actually has gotten quite good. But it had one problem over time that is well recognized, debated heavily, and has been an issue since the very first scanners came out, and that is, relative to the baseline, trying to figure out actual real anatomy and then how to compare, in this case, either above or below with the super-pixel analysis, has been burdensome and problematic. I think we've overcome that actually, and I'll touch on that at the end.

We have made a lot of strides. You know, there is good data, and we owe a lot to people in this audience about what strides we've made.

This is, first, from Zangwill et al., in reference to the OHTS database. And the conclusion here was that several HRT measurements,

including cup-to-disc ratio, cup volume, et cetera, at baseline were significant predictors. So this gets at that risk profiling, if you will. As well, the Moorfields Regression Analysis, which I showed earlier, being outside normal limits, also significantly predicated who was going to develop glaucoma.

Other strides we've made. Medeiros et al. talked about replacing this, and instead of looking at photographs, which I already highlighted have been problematic, you could actually replace HRT measurements into these risk calculators and they work quite well.

So we've made great strides. I don't think we predicted this 20 years ago. We didn't think that we'd be using HRT as a risk predictor.

However, the agreement breaks down at several points. And this is a paper that we did with a group in London a couple years ago, and I just pulled one slide from it to say that there isn't overall agreement. HRT doesn't totally agree with stereo photos. I don't think we should beat ourselves up about that. We have different sensitivity and specificity parameters for these. But it just points out that, while the middle group there, the Venn diagram, actually agrees, there are large groups that don't agree even between these two relatively simple technologies.

So why are we here today? I think we're here today to say is there a game changer here? Have we taken the next step with OCT? And I'd submit to you that we have. And I think the reason is multifactorial. One is the resolution of imaging, as we'll hear throughout the day, has gotten better

and better. But perhaps more importantly and highlighted by a couple papers by Burgoyne and Chauhan over the summer, and that is, we're now getting to the point where we can truly register these and have markers within the nerve that are stable markers that we can measure from. So we can measure up from -- they call it the Bruch's membrane opening. We can adjust for tilt. We can correct.

I think it's going to do several things. I think it's going to improve our structure function relationships. I think it's going to allow us to actually take the big step forward in terms of measuring progression, not just diagnosis, which I think is still problematic because of the variability of appearance, but we'll be able to more quickly tell.

So I think the next few hours will sort of reveal how this game changer is going to change what we do.

Thank you.

(Applause.)

DR. JAMPEL: Thank you, Jack.

The next speaker is Murray Fingeret, who will talk about GDx.

DR. FINGERET: Good morning. My talk is about the GDx retinal nerve fiber analyzer.

Here's a list of my disclosures pertinent to this talk.

And in many ways my talk is historical, about an instrument that for the most part has really fallen off in terms of use, but 15 years ago

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

was an important instrument, 10 years ago, in regards to the diagnosis of glaucoma.

The point being is that there are several ways that we evaluate the eye in regards to glaucoma. And in terms of imaging instrumentation, the original one was to look at the nerve fiber layer, and then Jack just talked about looking at the optic nerve. And as we talk about, for a good part of this day, OCT and newer technologies, this is a talk that is really a reference that goes back in history, in that glaucoma is a progressive optic neuropathy, and there's loss of the ganglion cells and their axons, and that's clinically manifested by thinning of the neuroretinal rim as well as loss of the nerve fiber layer. And it's the nerve fiber layer that this particular instrument is looking at and that scanning lays a polarimetry.

The GDx originally was an instrument that Laser Diagnostic had developed and Carl Zeiss has since marketed.

And as we look at the nerve fiber layer and the pattern as it goes towards the optic nerve, the instrument is going to evaluate or look for thinning of the nerve fiber layer.

Here's an example of an eye where we see a series of streets, the nerve fiber wedge defects, and in this case it's very easy to see. But there are many other examples where it's far more difficult to see.

So scanning laser polarimetry is based upon the principle that polarized light passing through the birefringent nerve fiber layer undergoes a

phase shift that's known as retardation, and this instrument is capable of detecting and measuring the size and depth of these nerve fiber layer defects. So it's just a diagram that shows how the light, as it goes into the eye, it becomes two polarized components, and that is going to be phase shifted as it reflects back and that the difference in the two rays coming back is going to correlate with the thickness of the nerve fiber layer.

One of the points about the GDx, and it's probably the point that we would use for OCT imaging and future technologies, is that they evolve. These are works in progress. The GDx underwent many changes from the original way it worked to the VCC, the variable corneal compensation, and then now to enhanced corneal compensation. And it's simply that there are different ways that they become picked up and measured and how the instrument is capable of fine-tuning to measure just what it wants to be. And in the case of the GDx, it evolved to be able to reduce any of the other components.

And then this became the printout. And the printout for the GDx, the top was simply a picture of the back of the eye, a false color image. The middle image was more of a gray scale or shadings, in which the brighter colors would represent thicker nerve fiber layer. And on the right side it's a duller blue around the optic nerve, indicating that the nerve fiber layer is thinner. And then on the bottom is pixels showing where there is some loss based on normative data. And then in the middle there's that so-called TSNIT

graph of showing the thickness of the nerve fiber layer around the optic nerve. There is evidence that polarimetry was successful and able to predict and recognize nerve fiber layer loss, as well as predict glaucomatous change.

So the strengths of the GDx were that it was capable of measuring the nerve fiber layer thickness that was associated with glaucomatous damage. We had a long history of use.

Weakness. Well, the instrument itself only measures the nerve fiber layer, it didn't evaluate the nerve or the macula, so it was only pertinent for glaucoma. And as we look at OCTs that are capable of multi-uses, that became a weakness. And there were atypical images that would come and that would affect the quality.

This is a paper that simply shows that, as the instrument evolved and went to a newer form of reducing the corneal component, the structure function relationships got better and better and better.

So nerve fiber layer imaging works, in that these imaging instruments, do improve over time.

Well, the last thing I want to just mention, you know, what's the agreement across platforms? How do the GDx results compare to OCT? And that becomes an important point because nerve fiber layer with the GDx was the first way, one of the first instruments or the first instrument to measure the nerve fiber layer, and now that's an important part of how OCT instruments evaluate the back of the eye.

And this is a case of just looking at the right optic nerve and the left optic nerve. A little hazy. The left eye has far more nerve fiber layer and larger cupping, nerve fiber layer loss, and you can see it right at the bottom right, where you see all of those pixels and the nerve fiber layer reduced.

An HRT showing this loss.

And here is a Cirrus. This is an example of an OCT also showing that that is being picked up in a similar amount.

And this is a paper that simply is looking at how now the structure function relationships were seen with both instruments. But perhaps, as we look at the OCT, it may even be doing at least as well and possibly even a better job at recognizing loss.

So nerve fiber layer assessment is an important method and the GDx was the start of which led to the -- is one of the ways of measuring the nerve fiber layer, which has become an important part of imaging instrumentation.

Thank you.

(Applause.)

DR. SINGH: Our next speaker is Joel Schuman, who will compare spectral domain and time domain OCT.

DR. SCHUMAN: Thank you very much for the invitation. And I think that this is a great meeting, and it's good that the American Glaucoma Society is partnering with the FDA in this process.

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

I'm showing here my disclosure. I do receive royalties for intellectual property related to OCT invention.

And what I'm going to talk about is the development of technology and about time domain and spectral domain OCT, similarities and differences.

Can we have the lights down a little bit up front? Is that possible? That would be no. Okay.

So let's go ahead with the video. I want to show you a brief video that was put together by the group involved in the invention of OCT, for the recent Champalimaud award.

(Video played.)

DR. SCHUMAN: Okay, can we go back to the slides, please?

All right. So that should give you some idea of the background of how OCT came to be.

And let me tell you a little bit about the technology itself. It's a Michelson-type interferometer. The time domain OCT, which was the first version of OCT, uses a moving mirror in order to tell you where in space a reflection from the eye is coming from. So the reference arm has a moving mirror, and that's how you know how far the light has traveled, the reflection in the interference pattern that you're receiving back.

With spectral domain OCT, instead of having a moving mirror, the mirror is stationary, but you're not measuring every pixel along the way.

You're measuring the wavelengths that are coming back in that A-scan all simultaneously, and those wavelengths tell you where the reflection is coming from in the interference pattern.

This is the schematic of time domain OCT, and you can see the moving mirror.

And this is a schematic of spectral domain OCT. No moving mirror, but it has a spectrometer and a line camera, in order to collect the information and then transform it into what we know is an OCT using Fourier transform.

There are two things that give OCT high resolution. One type of resolution that we talk about is transverse, or lateral resolution, and the other type is axial. And it's really the axial resolution where OCT has an advantage. And with commercial devices, you have axial resolution in about the 5  $\mu$  to 7  $\mu$  range, whereas the transverse resolution is about 20  $\mu$ , which is the size of the waist of the spot that's shining on the retina.

So time domain OCT has this moving mirror, a moving part, that is telling you where in space the reflection is coming from. And because of that, it's much slower than collecting all of the light and measuring the wavelengths to know where the information is coming from.

So Stratus, which was the only time domain OCT sold in the U.S., could do 400 A-scans per second. And it wasn't feasible to make 3D images using time domain OCT. But with spectral domain OCT, it is possible

to make three-dimensional images, and that changes the game.

So it's much, much faster than time domain OCT, and we're able to improve the reproducibility, the resolution, the things that we can do with the data that we can acquire, the ways that we can analyze them, and we can segment more accurately and register and orient the images more precisely.

So time domain OCT, 400 A-scans per second, it still exists. It still works. The discrimination between healthy and glaucomatous eyes using time domain OCT is no different than that discrimination using spectral domain OCT. And I'm sure that we'll be hearing about that today.

With spectral domain OCT, we're talking about much faster image acquisition. The resolution is on the 5  $\mu$  to 7  $\mu$  order. And there actually are commercial spectral domain OCTs that are faster, but not yet on the market, but approved by the FDA, up to 70,000 A-scans per second in the U.S. And then the next generation will go even faster than that and have higher resolution. And what we're looking at is much more information captured in a given amount of time.

And then there's swept source OCT, which we haven't talked about really as an entity. But swept source OCT goes back to just using a signal detector instead of a line camera, and it is using a sweeping laser that's going through the different wavelengths, in order for you to know where the reflection is coming from in space. So you can go very, very fast with swept

source, and there have been reports of over one million A-scans per second using swept source OCT.

So in the next five years, I expect that we'll see competition among the manufacturers, which is great for us because it will drive innovation forward and drive prices down. We'll be able to assess the structures of the eye like a pathologist would, except that it's noncontact and noninvasive, unlike pathology.

Software is continuing to advance, and we're learning how best to interpret the three-dimensional information that we're getting. And I think there are some talks about that relevant to glaucoma. And we'll be able to measure the subtle changes in pathology, addressing the issues that you heard about in the first section, with regard to early detection of disease. And I also expect that the cost will decrease as competition continues.

We can do optical biopsy with OCT. We're able to see the retina at the level of the cellular layer, and we can segment it. We can image in 3D, which allows us to register images and do follow-up in a longitudinal fashion much more precisely than we could before. And I think that we're going to see many clinical studies using OCT to define biomarkers for earlier diagnosis and tracking disease progression and response to therapy.

I want to thank the glaucoma imaging group at the University of Pittsburgh, as well as Jim Fujimoto and his group, for working with me on this technology.

Thank you very much.

(Applause.)

DR. JAMPEL: So our last speaker before the break will be Teresa Chen, who will speak to us about "Clinical Applications of SD-OCT."

DR. CHEN: So I'd like to thank the organizers of the meeting for the honor of speaking to you today.

When I first looked at the assigned title of my talk, I said I could talk about all the clinical applications of spectral domain OCT in detail. However, this is clearly not possible in 10 minutes. So instead I'm going to talk about just general concepts about imaging glaucoma, and hopefully this will be a springboard or transition to the other more detailed talks later in the day.

So perhaps another paradigm shift is spectral domain OCT, meaning, this is perhaps a transition from a more 2D, qualitative, subjective analysis of our patients to perhaps a more three-dimensional, quantitative, objective analysis of data in the clinic.

So the talk will be in two main parts. One is to look back at the literature, basically development of OCT and specifically how this relates to the clinic, and then looking forward, now that we have all of these spectral domain OCT machines, how this can be used for better in-clinical care of our patients.

So spectral domain OCT development, of course, cannot be

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

discussed without mentioning again the developers of OCT technology in general. And this was described already elegantly here by Joel.

Of course, time domain OCT, the basic setup is here. We have the SLD light source which shines light to the eye in a reference mirror, and this is processed by a photo detector to create the images that we do see in clinic.

So the commercially available machines that have basically dominated our clinical practice up until 2006, of course, have been time domain OCT machines, including the Stratus OCT-III. However, these technologies, as you all know, have lower resolution and slower acquisition speeds. So to address this issue and to better improve our clinical care, spectral domain OCT. And, again, similar output in terms of the light source, but different in how the images are processed or the data is collected. And this is, of course, with a spectrometer, which undergoes Fourier transform to create an image.

The early iterations of spectral domain OCT were developed by Dr. Wojtkowski. However, it was not until Johannes de Boer, who I've had the pleasure of collaborating with for over 10 years, that we have what's called video rate spectral domain OCT. And, of course, video rate is the key word because that's synonymous with 3D OCT.

So how do these imaging technologies fit, in terms of comparison with clinical exam? Of course, we know the gold standard clinical

exam is the eye exam with disc photo and visual field testing. And studies have showed that perhaps imaging modalities are better than a general eye doctor's exam. But on the other hand, they may not be better than a glaucoma specialist.

But let's say imaging technologies are similar to a glaucoma specialist. There are two potential advantages for imaging modalities, which, of course, is we have more objective data as well as quantitative data.

So clinical applications of all the imaging technologies. Of course, this is best summarized by the Academy's statement in a paper that was led by Shan Lin and, of course, Kuldev Singh and other co-authors, including Joel Schuman. And this is a literature review summarizing the technologies up until, again, 2006.

And this paper points out a couple of interesting points. One is that some of our glaucoma patients progress only by structure and some progress only by function. And this underscores the importance of our having structural imaging tests, because that may be the only indicator that our patient may be progressing in terms of their glaucoma. Another point again is that this data is quantitative and objective. And then, lastly, these imaging devices can be useful in clinical practice, especially for glaucoma diagnosis and determining disease progression.

So to briefly summarize some of the advantages of spectral domain OCT, we, of course, know that it heralds a transition from 2D to 3D

imaging, or, again, video rate imaging. A second advantage, of course, is the higher resolution. And looking at more experimental machines, the resolution can be down to  $2\text{ }\mu\text{ +/-}$ , compared to time domain OCT. And then the last benefit is perhaps better reproducibility.

Now, better reproducibility or less machine variability is important, because if my machine has less variability, I, as a clinician, can better determine if a change on that machine is due to machine variability versus actual real glaucoma disease progression.

So looking at the data of the coefficient of variation, we see that time domain OCT has larger coefficient of variation compared to spectral domain OCT.

Looking at a specific example of nerve fiber layer thickness, if you look at test/retest variability, this is clearly higher with time domain OCT. However, looking at even the best parameters, we see that spectral domain OCT has less machine variability.

So looking forward, now that we have all of these spectral domain OCT machines, how do we see if these new OCT parameters -- are they potentially useful for the clinic? And there are two things that we consider. One is, are these new parameters consistent with the current gold standard methods? And, two, is it consistent with known glaucoma pathophysiology? And, of course, there are other FDA speakers who are going to talk about validation in a more sophisticated way.

But simplistically speaking, if we have -- this is one specific example of a new OCT parameter developed by my colleague, Johannes de Boer. And here, if we look at this OCT parameter, we again first want to compare it with clinical gold standard methods. How does it compare with (1) disc photo and (2) visual field tests?

And so, for example, we see that there's no retinal rim parameter, that the neuroretinal rim thickness increases as cup/disc ratio decreases. So, of course, this makes sense, as it is consistent with gold standard clinical methods of evaluation.

Similarly, if you have an increase in neuroretinal rim thickness, this is associated with a more normal visual field test. Again, this is consistent with known gold standard clinical methods.

And then, secondly, we need to make sure that these new methods are consistent with known glaucoma pathophysiology. So, for example, if this machine tells me that my patient has thinning of the inferior nerve fiber layer in perimetric glaucoma, this better be associated with a corresponding superior visual field defect, again, instead of inferior visual field defect. So, again, consistent with known glaucoma pathophysiology. And, of course, good structure function correlation is the subject of many studies in the literature.

And then, lastly, looking at how spectral domain OCT can help in the clinic. Of course, other speakers will talk about improving glaucoma

diagnosis. Studies have shown that spectral domain OCT can be better in terms of diagnostic sensitivity compared to older imaging modalities.

Normative databases, of course, are very important because I, as a clinician, can better determine if a patient has glaucoma and so I have a better sense of what the normal variation is. And, of course, this is development of normative databases.

Spectral domain OCT can also evaluate not just the macula, the optic nerve, and then nerve fiber individually, but of course, perhaps a combination of these parameters can help us better diagnose glaucoma.

And then, lastly, as ophthalmologists, we don't want to be myoping and look at only structural tests. Our OCT data has to be examined in the context of other functional tests as well as our clinical exam.

So, second, longitudinal assessment of disease progression. Again, better reproducibility and less machine variability is important. As you all know, there are many longitudinal studies that are ongoing, and hopefully data from these will be available in the next few years. So when these studies are out, hopefully this will better define disease progression in our patients with imaging.

And then, lastly, risk stratification. What does this mean for my patient? If I have an abnormal OCT value, my patient wants to know, is this associated with good or poor prognosis?

So looking at older imaging technologies, and new ones, we

know that in post hoc analysis of prospective studies, that certain baseline imaging parameters in suspect patients, or ocular hypertension patients, can be associated with the future development of glaucoma. Also, of course, imaging devices help us better define pre-perimetric glaucoma.

And then, lastly, imaging does provide, again, objective and quantitative data. And, of course, objective and quantitative data can better improve clinical trial endpoints as well as hopefully help us to better determine if new treatment strategies are effective.

So, in summary, many exciting developments in imaging in the past decade or two, and then looking forward, hopefully we'll be able to better understand how the existing spectral domain OCT machines can improve our clinical care.

Thank you.

(Applause.)

DR. SINGH: So we'll have a short break now, and we'll resume at 9:45.

(Off the record.)

(On the record.)

DR. GREENFIELD: Well, once again I'd like to welcome everybody to Washington and welcome also -- with no disrespect to any of our prior speakers, welcome to the meat of the program. This session will fundamentally deal with some of the complexities in OCT clinical science as

well as OCT regulatory science.

My name is David Greenfield from the University of Miami. I'm joined by my co-moderator, Gadi Wollstein, from the University of Pittsburgh, who will introduce this session and then choose the speakers.

DR. WOLLSTEIN: Thank you, David.

This session will focus on the characterization of OCT as a measurement tool. The clinical application of an imaging device, in general, is markedly dependent on the measurement accuracy and test/retest variability. And as we are going to hear in this session, there are numerous factors that affect these properties.

And the outstanding speakers that we have in this session will review source of OCT measurements, artifacts, and examine the agreement, precision, and variability of OCT parameters.

The first speakers will address factors that impact image quality and measurement. The first one will be Dr. Vizzeri from the University of Texas in Galveston -- I actually remembered the full affiliation -- that will address the patient-dependent and the operator-dependent factors. Followed by Dr. Hood from the University of Columbia, that will address the instrument-dependent factors.

DR. VIZZERI: Thank you and good morning. I want to thank AGS and FDA for inviting me today.

These are my financial disclosures.

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

So today I'll be discussing factors that impact image quality measurements, and these factors can be divided into patient-dependent factors, operator-dependent factors, and instrument-dependent factors. And I'll briefly touch on some of the patient-dependent factors such as pupil size or media opacity, operator-dependent factors such as centration and alignment, while instrument-dependent factors will be the topic of the next presentation.

So, first of all, why should we bother knowing about all of these factors? So let me just make an analogy with space exploration. So if you're trying to know more about a planet, certainly you would want to know the weather on the planet and the conditions on the surface or the temperature, and you also would like to make sure that your operators get there safely, in good health. You're concerned, obviously, about getting there in one piece.

All of these factors, unfortunately, are linked together. And what makes it very challenging for us in our field is that we don't just have one machine. We are now dealing with multiple machines.

In the past, as mentioned before, there was one time domain OCT. The newest version was the Stratus OCT. And now we're studying multiple spectral domain OCTs. These are all different machines, they're all available on the market, and they have different hardware and different software. It's possible that even factors affecting variability may differ from one machine to the other. Most of the studies are done with Cirrus. I

mentioned the ones from Stratus OCT. Cirrus, incidentally, is also the one spectral domain OCT that I use in my practice.

So let's move to the first patient-dependent factor that I will discuss. It's pupil size. And although a small pupil size might affect time domain OCT measurements, from reviewing the literature it appears pupil diameter does not significantly affect spectral domain OCT results. So we can say with some confidence that OCT testing can probably be performed without pupil dilation listing in most of our patients.

Let's move to a more interesting patient-dependent factor, media opacity. So, for example, the presence of a cataract. Unfortunately, media opacity does decrease the signal-to-noise ratio. And with time domain OCT, well, the machine may fail to identify the upper and lower borders of the RNFL on low signal-to-noise ratio scans. In general, a lower signal strength may result in lower RNFL DCAs estimates.

Here's a nice graph showing on the X-axis the clock-hours and the Y-axis the RNFL thickness. In upper right, you have different signal strength values. And you can see the lower the signal strength and the lower the RNFL thickness across all clock-hours.

So if we're measuring RNFL thickness with time domain OCT, it is important to aim for a signal strength of at least seven. And this also seems to be the case for Cirrus OCT. OCT scans with signal strength equal or less than six should probably be discarded.

Let's look at another probably important patient-dependent factor, anatomic variations. And some of these anatomic variations that were suggested are optic nerve tilt or peripapillary atrophy. Reviewing the literature, it seems like more evidence is needed to evaluate the impact of these anatomical features. But it's important to note that optic nerve tilt and peripapillary atrophy often occur in myopic eyes. And we know from many studies that refractive errors, axial length, these may affect time domain and spectral domain OCT measurements and classification results, as probably will be discussed later on.

Normative databases currently do not include eyes with significant refractive errors. So in my opinion, these results should be interpreted very cautiously.

Let's move on to operator-dependent factors. And by far, what I think is the most important operator-dependent factor, at least with time domain OCT, is centration. And what I mean by that is the position of the circle scan around the optic nerve.

Time domain OCT RNFL assessment relies on the operator to consistently center the circle scan around the optic disc during each follow-up examination, and there's no automated registration of OCT scans available.

So here's an example of improper scan alignment. You can see in the second and third scan, when the scan is either too close to the superior disc margin or too close to the inferior disc margin, you cannot measure

changes in the sector RNFL thickness measurements.

Not only that, but we have shown that the other RNFL thickness might also change. In particular, with horizontal scan shifts, the average RNFL thickness increases with temporal shifts and decreases with nasal shifts. And this also has been demonstrated longitudinally.

So this issue with scan centration potentially may limit the ability of Stratus OCT to detect true glaucomatous change over time.

Thankfully, most spectral domain OCT devices now employ automated algorithms to ensure the circle scan is centered on the optic nerve at baseline, and also scan registration features, to consistently place the scan in the same location at follow-up visits. Spectralis even uses an eye tracking system for this purpose.

So I want to show you, as an example, what Cirrus is able to do now. First of all, it identifies the presence of the disc on the en face OCT image. Second, it defines the disc margins. After the disc margins are outlined, what Cirrus does, it places the optic disc center automatically. And, finally, it places the circle scan evenly around the optic nerve. So the process now is fully automated and operator independent.

Let's look at other possible operator-dependent factors. These mainly deal with the instructions that you give to the patient. I tell patients that are undergoing an OCT scan that it is very important to keep your head straight. This is because head tilt can have an impact on RNFL thickness

measurements, as shown nicely in this study.

Second, I tell them to maintain fixation and do not blink. This is because we're seeing numerous motion artifacts on a lot of these OCT scans, particularly with spectral domain OCT, and it's possible that these motion artifacts may affect measurement variability. There is one recent study that suggests that these motion artifacts might not be a major issue.

Finally, how to assess reliability and quality of a scan. Well, as a clinician, what I have in front of me is a printout. And here's an example of the Cirrus printout. So what I recommend is not just to look at the printout for glaucoma detection and progression, but also for quality purposes. What I mean by that is, at first, you should always look at the date of birth and the sex to ensure that the normative database correctly applies to the patient that you're examining.

Secondly, you should look at the signal strength. The lower the signal strength, the lower the RNFL thickness. So I look to see if the signal strength is at least seven in both eyes.

And then you have the RNFL thickness maps. These maps are very important as far as quality goes, because I can check to see if the disc is correctly placed in the middle of the OCT map. I look to see if there are any motion artifacts, any missing data, signs of algorithm failure. I also look to see if the disc margins are correctly identified and finally if the circle scan is correctly placed around the center of the disc. I also look at the sector and

clock-hour RNFL thickness to look for major discrepancies in the measurements.

So moving on to the FDA questions that I thought were very thoughtful, with regard to this subject, are there additional sources of variability that should be evaluated? In my opinion, yes, there may be additional sources of variability. Some have actually already been pointed out, motion artifacts, intraocular pressure changes, even heartbeat. But the question, in my opinion, is which sources of variability are of clinical relevance for glaucoma? It's possible that several of these sources of variability may not be clinically relevant.

The second question was, should studies be designed to allow separate estimates of variability due to the device and due to the operator, or is it acceptable for these sources of variability to be confounded? Ideally, in my opinion, it should be nice to have studies designed to look separately at these issues. But remember what I showed before. These spectral domain OCTs are becoming more operator independent. They're user-friendly machines, and so operator-dependent variability might be less of an issue with these machines.

So, in summary, identifying factors impacting image quality and measurements is essential to optimize precision and allow for the detection of progression of glaucoma. And even if we cannot completely eliminate these factors, at least knowledge of these factors may help establish a hybrid

size selection criteria.

With regard to spectral domain OCT technology, this technology has the potential to overcome many limitations of time domain OCT. These machines are becoming more operator independent. However, in my opinion, further studies are necessary to better characterize factors affecting spectral domain OCT variability.

So going back to the analogy made with Mars exploration. Well, what NASA has done in the past years, and is still doing, it has essentially removed the operator-dependent component. And these machines, although they're not perfect, are now doing an excellent job exploring the surface of the planet.

Thank you.

(Applause.)

DR. WOLLSTEIN: Our next speaker is, as I mentioned before, Dr. Hood from the University of Columbia, that will discuss the "Instrument dependent factors" that affect the measurements.

DR. HOOD: Here are my disclosures. And also, in the spirit of full disclosure, let me say that I am not an expert on either image segmentation or image registration. I do not understand the technical details, but I was told that's not important in terms of what I'm supposed to be doing here.

So what I'm going to do is I'm going to make some points about

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

image segmentation and then raise two questions about image registration. So let's start with image segmentation first.

The algorithm used to segment the layers is probably the -- I say "a," but I think it's "the" key factor determining whether the results from different machines will be compatible. And we learned this actually in a study we did three years ago, in which we compared results from the Stratus to results -- at that time it was a fairly new spectral domain, which I'm calling frequency domain, machine by Topcon. And what we were looking at was the average retinal nerve fiber layer thickness for peripapillary circle scans around the disc. And what we found was that the agreement across machines was good. In fact, it was very good.

So for this group of control suspects and glaucoma patients, the agreement mean difference was less than a  $.10\ \mu$ . And, in fact, you can see it here, where we plotted the spectral domain versus the Stratus, and you can see the points clustering around the slope of one. If you present it as a Bland-Altman plot, they cluster around the line of zero. There's a slight slope to this, and that's because some of the extreme glaucoma patients here are showing a greater thickness on the frequency domain than they are on the Stratus.

So let's take a closer look at one of those patients to see one of the reasons why.

So here is the peripapillary thickness. That's the Stratus.

There's your frequency domain, right? And one of the differences that's causing the thickness difference is that the Stratus was less influenced by blood vessels when the retinal nerve fiber layer is very thin.

So let's blow this up. Here's a blood vessel here, and you can see that in the Stratus here, the blood vessel is being ignored by and large, whereas in the frequency domain, it is being included.

So algorithms differ. And the reasons in this particular case is this very early algorithm in the Topcon machine had very little spatial averaging, whereas the Stratus has a lot of spatial averaging.

All right. Now, the second point I want to make is that all of these segmentation algorithms involve tradeoffs. There's typically not a best answer because you're making some kind of a tradeoff here. I'm not saying one is right and one is wrong. It's going to depend on the circumstances.

So, for example, here's a study we did in which we were looking at between subject variability in both OCT and visual fields. And I just want to do the OCT part. And this was all done with Stratus, so we're only looking at one machine here.

So what I want to point out is when the vessels -- due to spatial averaging, when the vessels are included near other vessels or with the retinal nerve fiber layer tissue around it, the algorithm is going to include it when they're isolated, because there's a lot of spatial averaging that's not included. Okay. So because of spatial averaging, the blood vessels are

omitted when seen in isolation, but included when close together or combined with the presence. Now, this is not an argument for more or less spatial averaging because, again, it's a tradeoff.

So here's an example of a frequency domain machine. Because of the spatial averaging, now you're picking up this shadow and you're not including the blood vessel. So there's no right answer. It's a matter of tradeoffs.

Okay, the third point, algorithms differ in lots of ways. So you've got preprocessing going on. You've got the core algorithms that are doing your segmentation. I just spoke about smoothing. You've got the order in which borders are identified, which can affect results, as well as special methods to deal with special structures such as blood vessels and raphe in fovea.

Dimensions can even differ in their definition of borders or layers. So, for example, I think most of you are familiar that there's a variation in what's meant by retinal nerve fiber layer thickness, depending on which of these layers you use for your distal border of the retina.

Five, scan quality. And I guess the previous speaker touched on this. And thus, scan protocol will affect segmentation.

So there are a number of nice studies in the literature. I chose this one because I like the pictures, in which they're showing that the image quality here of a scan for the same patient, taken about at the same time in

the same section, right, the thickness will be different depending on the scan quality.

Point six, all algorithms make mistakes. So in addition to all the things that the previous speaker said, you should look at when you're looking at a report, you ought to also look at how well the algorithm is doing. Okay, sometimes mistakes are very obvious, right? Sometimes they're more subtle. But all algorithms are going to have mistakes.

Okay. So how do we go about validating algorithms? I'm going to mention three possible methods. I'm sure other speakers will suggest other methods.

We've done a lot with human manual segmentation as our gold standard. And so I can tell you, based on our experience, that manual segmentation produces extremely good within and between operator or segmented reliability. In fact, it's probably the best way to -- I'm using predicate here, I think, in a different way than FDA is using it. It's probably the best way to validate segmentation.

But training and a manual are necessary. That is, you just can't have somebody who has looked at all of these scans do the manual segmentation and expect them to do the same thing next week or do the same thing that you might do. But with a little training and with a manual, you can get extraordinarily good results.

A second approach is a little more subtle, so I'm going to have

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

give you some background. You can use structure function analysis, I realized recently. And here's what I mean by that. Let me just tell you about a study first.

Here's a paper that we published in *Archives* at the end of last year in which we compared the 10-2 local sensitivity loss to local fitting of retinal ganglion cell in a plexiform layer thickness. So we wanted to know how this was related to local sensitivity loss, and we used an automated algorithm, but we hand-corrected it for this study because we wanted the best possible data.

So here's a sample patient. You can see there's a thinning. There's a field defect here and a thinning here. So we're going to flip this so that we have field view for both of them. All right. Now what we want to look at is we want to compare for fixed eccentricities, local sensitivity loss, and local thickness. All right. So let's take these points here as an example, these four points.

All right. Now, the light from those four points fall right here. All right. So what we can do for each point, we can plot the sensitivity loss on the X-axis and the thickness of the retinal ganglion cell layer on the Y-axis. Clear? That's the patients, that's the controls, and we get a messy looking thing with a lot of variability. Much of that variability, but not all, is due to the fact that we did the wrong comparison. So stay with me here. Let me remind you that although the light is falling here, the ganglion cells of interest

are out here, because light falling on these receptors stimulate ganglion cells out there because they're pushing the cell light.

Now, fortunately there's this nice study from Christine Curcio's lab, Drasdo et al., in which they traced in postmortem histological tissue the root from the receptors to the associated ganglion cell and then presented that evidence in a mathematical formulation, so we're able to adjust this and say that on average, based on this study, the light here stimulates ganglion cells out there.

All right. Now what can we do? Well, what we can now do is correct for that. So now we're plotting this location, and you can see, although there's still variability, you get a much nicer relationship.

So here's how it can be used for validating algorithms. This is the corrected algorithm that was hand-corrected. You can compare that now to an uncorrected algorithm and see how well the algorithm is doing.

However, I want to warn you, this is a very, very sensitive test of the algorithm. That is, this local structure versus function analysis might be of use in validating algorithms, but it's undoubtedly much, much stricter than you need for clinical validation.

So what do you do for clinical validation? Well, I don't know too many ways to get around sensitivity and specificity analysis. So you could evaluate a particular algorithm and compare it to a manually corrected version of it. Or you could compare algorithms, as long as you have the same

group of patients. Okay. So for clinical validation, sensitivity and specificity analysis may be a best approach.

However, here's another warning, and this is another subtle point. The algorithm that works best for detection may not work best for progression. And we can discuss this more later, if you want to know why I'm saying that.

Okay. So what about image registration? There are different types of image registration. I know even less about this. There are different types of image registration. There's registration that goes on when you want to average multiple B-scans. There's registration that goes on when you want to realign B-scans within a cube scan so that you correct for motion artifacts. There's registration that goes on when you want to look at a patient over time and compare scan at time one to scan at time two. And then there's also registration algorithms for comparing scans to other kinds of images such as a fundus or a fluorescent photo.

So point number 10. As with segmentation algorithms, there are tradeoffs. Thus, there may be better methods, but probably no right or best method.

Finally, I want to end with two questions about image registration that I have. One is, how much detail should manufacturers provide to users? At the moment we're being told almost nothing. Is that okay? And, second, should we worry about these factors, or is the proof in

the segmenting? That is, if we validate the segmenting, are we happy enough with whatever is going on for the registration?

Thank you.

(Applause.)

DR. WOLLSTEIN: Thank you.

And I want to invite the next speaker, Dr. Hilmantel from the FDA, that will define for us the terminology and the concept used for measurement validation.

DR. HILMANTEL: Good morning. I want to thank Lee Kramm and other FDA reviewers for helping me with this talk. Dr. Kramm is no longer with FDA, but he contributed significantly.

I'm with FDA. I don't have any financial interests at all.

Okay, the purpose of this talk is only to define basic concepts to make sure that everyone is on the same page. I'm going to talk about these various topics here. And the regulatory route, as Dr. Eydelman discussed, is the 510(k), where we have to establish the substantial equivalence between a legally marketed device and a new device.

These are the various things that FDA looks at to validate the measurement techniques. We go through bench testing, segmentation agreement, agreement of precision in clinical studies, also.

In the bench testing, we look at resolution testing, both axial and lateral, validation of thickness measurements and validation of lateral

measurements. The bench testing would be greatly facilitated if we had a retinal model or phantom that was accepted by everybody. Unfortunately, there's no widely accepted phantom at this time. Dr. Dan Hammer's group in FDA, at OSEL, is working to establish a standard phantom that will greatly facilitate our development here.

And here's an example of the phantom that the OSEL group has developed. On the left you see the OSEL phantom and on the right the normal retina. They're very similar.

And Dr. Hammer is also working on adaptive optics techniques that will soon be applied to OCT-type devices.

The first thing FDA looks at is segmentation agreement. In this type of a study, images are supplied by the sponsor, and the trained experts have to demarcate the different layers. Measurements are made on the thicknesses of the layers, and these are compared to the thicknesses that are produced by the machine, and they have to agree.

And then we go to the clinical studies. Both the agreement and precision studies are involved. Agreement studies compare the measurements made by the new device to the predicate device, and the precision studies compare the repeatability and reproducibility.

In recruiting subjects, generally, subjects that are recruited are representative of the intended range of measurements and the pathologies that are being studied. Generally, there are three groups of eyes that are

used in these types of studies, the normal eyes that are free from disease, eyes with retinal disease or various types such as macular degeneration, diabetic macular edema and so on, and then there are eyes with glaucoma with varying severity of disease.

In the agreement studies, we characterize the difference between the measurements from the new device and the comparator. Generally, the type of study that's provided is outlined here. There's one eye per subject, one measurement per eye on both the new machine and the predicate device. The order of testing is randomized, and subjects are recruited from multiple clinical sites. Studies are performed separately on each group of patients, the normals, the retinal disease group, and the glaucoma group, and analyses are also separately performed on the three different groups. And the difference between the two device measurements are summarized in descriptive statistics.

FDA is also provided with plots to help clarify the comparative performance of the devices, and regression analyses are performed.

This is an example of difference plots. And here on the Y-axis, the difference between the two devices is on the Y-axis and the predicate device measurement is on the X-axis. In this type of a plot we expect to see a uniform distribution of points. But in this example here, you see that there's a trend downward. So for this type of a performance, we would expect some sort of explanation from the sponsor.

And in this graph, the Y-axis is the measurement from the new device, in this case, for retinal nerve fiber layer thickness, and the X-axis is the retinal nerve fiber layer thickness measured by the predicate device, and the graph displays both the regression line and the line of  $y=x$ , which would be perfect agreement of the two devices.

This is an example of the type of labeling that you would see in an FDA device, showing the results of the agreement study. And this table would be provided for each patient group, the normal group, the retinal disease group, and the glaucoma group.

So in this type of table, each row is one specific parameter, such as retinal nerve fiber layer thickness and the temporal quadrant, the superior and so on. And then you would have optic nerve head parameters such as C-D or rim volume or cup volume. And then the columns are the various outcomes, the average measurement for the new device, the average for the old, the average of the differences between the two devices, the confidence interval, and the limits of agreement.

The limits of agreement, for those who are interested in this, is supposed to characterize where 95% of the points fall, roughly. So for each eye, the difference between the two devices is calculated, and the mean of the difference is calculated, and the limits of agreement is the mean plus or minus two standard deviations of the differences.

So the questions that FDA are putting out for discussion are,

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

are there more clinically useful ways to describe the agreement results in the user's manual? For example, the percent difference versus the differences in microns. What difference in agreement, if any, between the new device and the predicate device would be of clinical concern? For example, if they differ by more than 10%, is that of concern?

Now we go on to the precision studies that FDA asks for. These characterize the repeatability and reproducibility of the new device compared to the old, and repeated measurements are made for a given eye on a given device and also for several devices of the same model. The patient is removed from the headrest and repositioned between measurements.

And this is an example of the study design sort of graphically. A study is performed for each patient category separately, the normal, retinal disease, and glaucoma groups. And the above study design is done for both the new OCT and repeated for the predicate OCT. The sources of variability are obviously many. As discussed, there are patient realignment factors, the internal machine variability, there are operator factors, variability from device to device, and in the study design just outlined, the operator factors and the device-to-device factors cannot be separated. There are also other factors which are somewhat unknown, day-to-day variations in the structural features, the thicknesses of the different layers, and so on.

A question for discussion is, considering the increased burden due to the greater number of measurements that would be needed, are there

additional sources of variability that should be evaluated in precision studies?

For example, day-to-day variability in glaucoma-related measurements.

And (b) should studies be designed to allow separate estimates of variability due to the device and due to the operator? Or is it acceptable for these sources of variability to be confounded?

This is just something that's of general interest to statisticians.

And here's an example of labeling results for a precision-type study.

So, again, a complete table is provided for normal eyes, for retinal disease eyes, and for glaucomatous eyes. So it's the same format as the agreement-type study results. In this case, the rows would show the repeatability standard deviation, the percent coefficient of variance, and the reproducibility standard deviation on the corresponding CV.

Another question for discussion is, as advances in technology continue to be made, we expect the precision to improve. What magnitude of difference, as described as a ratio of standard deviations between the precision of a new device and the predicate, would cause clinical concern? For example, would it be acceptable if the precision of the new device were worse than that of the predicate by a factor of two, three, or more?

And this is just a summation of our questions. So we have two questions pertaining to agreement studies, and there are two questions, which I just went through, pertaining to precision studies.

Thank you for your attention. I'm sorry my talk got out of order a little bit here.

(Applause.)

DR. WOLLSTEIN: Okay. So we have a slight change in the order of the presentations, and Dr. Hilmantel will continue with his second presentation right now.

DR. HILMANTEL: Well, I guess I have the right notes for this one here. This talk was supposed to be the first one, and this is just supposed to kind of give basic definitions of the concepts. This is very simple.

Just to make sure everyone's on the same page, again, Lee Kramm was heavily involved in this work also.

I have no financial interest in any of these devices or related areas.

Measurement validation is the process of characterizing aspects associated with device performance and measurement. It includes precision, bias, accuracy, and agreement. And this process is assessed through both bench studies and clinical studies.

Precision is the closest agreement between replicate measurements under specified testing conditions. The key words here are "under specified conditions." Repeatability is precision that's assessed under conditions that vary only minimally, and reproducibility is precision assessed under conditions that vary close to maximally.

Quantitatively, we actually express the imprecision of the measurement using the standard deviation or percent coefficient of variation, which is just the standard deviation over the mean value times 100. Thus, there's a repeatability standard deviation and a larger reproducibility standard deviation.

Bias just tells you how big the difference is between the measured value and the true value on average. Bias may not be constant across the range of possible results, and it may be different for healthy subjects versus subjects with disease.

The accuracy of the measurement varies its distinct concept from diagnostic or clinical accuracy of a test. Diagnostic accuracy is characterized by measures such as sensitivity or specificity. They characterize how closely the test results are associated with the disease of interest. Diagnostic accuracy will be discussed in a later session.

And this figure here just -- I think a lot of people are familiar with this. This graphically shows the concepts of bias and imprecision. The little bullet shots, the distance between the individual bullet shot at the center of the target is a measure of accuracy or it's analogous to accuracy. Bias has to do with the distance between the centroid of the bullet shots in the center of the target, and the imprecision has to do with the spread of the bullet shots. So you can see that you can get various combinations of high bias-high imprecision or low bias and high imprecision and whatnot.

Agreement of measurements has to do with how measurements from one device, how close they are to those of a device that's for measuring a similar parameter. It's often characterized by the mean and standard deviation of paired differences.

Precision bias and agreement measures can vary across the measurement scale and may be different between healthy and diseased eyes, and they may depend on image quality.

The terminology defined here will be used in the FDA presentations that follow, and in my previous presentation.

(Laughter.)

DR. HILMANTEL: Thanks for your attention.

(Applause.)

DR. WOLLSTEIN: Thank you.

And our next speaker is Dr. Girkin from the University of Alabama, that will discuss the agreement and precision of OCT measurement in different scanning locations.

DR. GIRKIN: Thank you. I'm going to kind of take off from that excellent talk by Dr. Hilmantel to look at the literature on agreement and precision of the OCTs.

I've worked on the normative databases for Optovue and Carl Zeiss and received some instruments from them five years ago or so, and we just started with a normative database study with Heidelberg on the

Spectralis.

So this is kind of a repeat. As Dr. Hilmantel discussed, repeatability is, within one session, a measurement of repeat measurements. Reproducibility is across sessions on different days, allowing for the maximal variance that probably reflects more real-world use of the instrument. This is not really the terminology used in the literature. It is the more appropriate terminology. But we often talk about this as inter-visit variation and intra-visit variation, as the way the literature used it.

In talking to the meeting organizers, we restricted this to the Spectralis, the Cirrus, and the RTVue, as those are the largest, most utilized instruments on the market. I also restricted -- tried to just use coefficient of variability when available, and we'll talk about why, because it allows for better comparison across the devices and then looked at agreement across the Spectralis, Cirrus, and RTVue. And also I included how the published literature agrees with the registration data.

While all of this is important, it is probably the boringest talk I'm going to ever give --

(Laughter.)

DR. GIRKIN: -- because a lot of literature is a lot redundancy.

So a coefficient of variation. The reason it's a good measure to compare across instruments is that it controls for variation of standard deviation with the mean. So it's basically the ratio of standard deviation to

mean. It's a dimensionless number, so it's unit free and allows that comparison. You can only use it with ratio scales. It can be used to develop confidence intervals for a unit level within an individual instrument. And there's one of the disc -- oh, it's not as good if there's scale bias. There are better measures if there's scale bias, which there are in some of these instruments.

So we look at the optic nerve. We'll just run through these papers really quickly. Gonzalez-Garcia looked at repeatability using the RTVue-100, using the standard 4 x 4 raster scan and the 12 mm radial scans -- the 12 3.4 mm radial scans. In this study, the contour lines were drawn manually. Three scans were done in a single visit. And they also looked at retinal nerve fiber layer, which had much better reproducibility.

The coefficient of variability, you'll see, like a disc area, is very good. They were drawn manually, so that's no huge surprise. But for some of the disc measurements, they were very, very high. Cup/disc ratio had 22.2 coefficient of variability, which is pretty high. Rim area was really low. And that disagrees significantly with the registration data, and I'm not sure how to explain that.

If you look at the repeatability and reproducibility of the RTVue instrument in 12 normal subjects and 12 controls, you'll see vertical cup/disc ratio with a very low coefficient of variability in repeatability and a pretty decent coefficient of variability, 6.6, in reproducibility.

Moving on to the OCT, the most recent and probably the largest study now is the Mwanza study from out of Bascom Palmer. It's looking at repeatability and reproducibility of optic disc analysis, but also looked at RNFL. And we'll review that later.

In glaucoma patients, they used the 200 x 200 optic disc cube scan centered on the nerve. They did five visits, five scans each, using different examiners. You can see the coefficient of variability for the optic nerve parameters in decreasing order and increasing order. Cup/disc ratio is the most reproducible parameter, a quite low coefficient of variability both for repeatability and reproducibility at 1.1.

They calculated a tolerance limit in microns, of 0.2  $\mu$ , between -- I'm sorry -- 0.2 in CDR between exams, as the instrument being able to detect a significant change, defined by one standard deviation times the square root of twice the in-subject variance. And we can talk about what the best measures are for tolerance. I think that's somewhat subjective, and there's some debate about it.

If we look at the registration data, the Cirrus -- this is in glaucoma -- the coefficient of variability was a little bit higher, 4.7% for cup/disc ratio.

So, in summary, just of optic nerve parameters in general, precision of disc area is good, but appears better, obviously, with manual outlining of the disc, which may be impractical for wide-scale clinical use.

There's a disagreement between RTVue registration data and what the literature shows, with cup/disc ratio performing worse than the registration data and rim area performing better. I'm not sure why that is. The Cirrus reproducibility is better in the clinical literature than the registration data. And the best reproducibility seen so far with the optic nerve head is the Cirrus cup/disc ratio with a coefficient of variability of 1.1.

So looking at the retinal nerve fiber layer. Again, the same paper, the Gonzalez-Garcia paper out of UCSD, looked at repeatability in normals and glaucomas, using the NHM4 scan with the RTVue-100 instrument, with 60 healthy control subjects and 38 glaucoma patients, and found quite a nice inter-session coefficient of variability in the glaucomas. That should be reproducibility and repeatability in the glaucomas of 1.9% on average for global RNFL. The sectors ranged 2.9% to 4.7%. In the normals, 1.5, a little smaller coefficient of variability, which you'll see across all of these studies, with a range of 2.7% to 3.9%.

Also, what you see across the studies is the nasal and temporal portions of the disc always had the highest coefficient of variability, which is good for us, since most the damage occurs superiorly and inferiorly.

So the registration data of the RNFL scans with the RTVue show somewhat similar reproducibility and repeatability. Reproducibility coefficients for RNFLs and normals is 2% and amongst glaucomas was .6. This is what you would expect.

Garas and colleagues also examined the effect of disease severity using the RTVue, and they grouped patients into normal in one group, mean defects between 6 and 12, and in Group 3 mean defects above 15. I'm not sure what happened to the patients who were between 12 and 15, but maybe they weren't there. I don't know. But the coefficient of reproducibility for the retinal nerve fiber layer -- these are just the global measures -- increased, as you would expect, with disease severity, and that was significant.

Moving on to the Cirrus, Hong and colleagues in 2010, looked at a repeatability study using the optic disc cube scan, three scans in a single visit, looking at 30 normal eyes. They had 52 total subjects, and they excluded 22 due to low signal and blinks. The coefficient of variability, globally, was 2.4%. And it's globally similar to the Stratus, but regionally it was superior to the Stratus, the predicate device.

Garcia-Martin and colleagues in 2011 also did a repeatability and reproducibility study in normal subjects, using a 200 x 200 optic disc cube scan, two to three scans each. Each exam had a different examiner. They used 72 normals. Their inter-session COV was .14% -- intra-session, which is repeatability. Reproducibility, inter-session, they didn't present the COV but had a mean difference of 2.7  $\mu$  and presented some Bland-Altman plots between the first and second visits, showing that variation really didn't change with variations in the mean nerve fiber layer thickness.

So now the Mwanza study, if I'm pronouncing that right, looked at -- in addition to looking at the optic disc cube, looked at repeatability and reproducibility in the same glaucoma patients, looking at RNFL. And, again, five scans, five visits, 55 subjects, a fairly large study, and showed the inter-session COV with the Cirrus instrument. For global RNFL it was 1.9% and intra-visit was 2.7%, with a tolerance limit of 3.89  $\mu$  for the global retinal nerve fiber layer.

Lastly, Wu and colleagues, using the Spectralis, looked at -- just recently looked at repeatability in normal subjects and in glaucomas, using the 3.5 mm ring scan with and without eye tracking, and they found, similarly, you had a high coefficient of variability in glaucomas than normals, at a coefficient of variability of 1.5 in the normals and 1.7 in the glaucoma patients. They found no relationship between the mean retinal thickness and standard deviation, as the other studies have shown, in either group.

Actually I made a mistake. This is A and B. Normal is A with eye tracking, and the glaucoma was supposed to be without eye tracking. So they're not normal and glaucomas. The lower coefficient of variability is seen with the eye tracking engaged between exams, and the higher coefficient of variability is without the eye tracking engaged.

Langenegger -- I knew I was going to butcher that name -- in 2011 also did a repeatability study in normals and glaucomas, using the 3.5 mm scan. Again, a very similar finding, Groups A and B, with a higher

coefficient of variability with using the eye tracking availability with the Spectralis. The Spectralis in these studies did use 15 scans per slice, but then they disengaged eye tracking for the Group B subjects between the session visits.

Now, this is the difference in coefficient of variability with and without eye tracking, so that eye tracking helped the glaucoma patients much more than it helped normal eyes. So the eye tracking improved the coefficient of variability in the glaucoma eyes. This is the delta COV between the two tests.

Looking at the registration data, the Spectralis, the coefficient of variability, repeatability was .7% for foveal thickness and for retinal nerve fiber layer thickness globally. This had the lowest reproducibility in its registration data as well. For foveal thickness, which repeatability is a little bit higher, as expected, 1.1% and 1.4%. Now, this equates to a confidence limit of between 1  $\mu$  and 2.5  $\mu$ .

Probably most importantly is what is the repeatability across these commonly used instruments. And there are a few studies out now. Pierro, in *IOVS*, has recently published a study of 38 right eyes of 38 volunteers. Each eye was scanned twice by two operators within the same day. So this is more of a repeatability than a reproducibility study. They tortured these patients with seven devices through the day, and they're all listed here. I'm just going to show Cirrus, RTVue, and Spectralis, and they

used the standard scanning parameters.

Looking at RNFL reproducibility, the highest precision was with the Spectralis, with an average -- looking at average nerve fiber layer with an average coefficient of variability of 1.65. The Cirrus was in the middle with a coefficient of variability of 2.2. And the Optovue was the lowest with a coefficient of variability of 3.3. All low, but there is certainly some difference in precision across the OCT devices.

Here's intra-operator variability showing variations between operators, and intra-operator variability is lower, obviously, than between two operators. So intra-operator variability is lower and better with each device.

Leite and colleagues used the DIGS database to look at RNFL agreement across the RTVue, Cirrus, and the Spectralis as well, and found similar measurements, but a little thicker measurements with the RTVue than the Spectralis, and the Cirrus had the thinnest measurements. And I'll show in another study the correction factors for this.

The Bland-Altman plot showed that this is not uniform across the average difference between these devices, and there are some fixed proportional bias in comparing these measurements.

The Pierro study also shows the measurement of this fixed proportional bias. They used structural equation modeling to develop calibration equations, and you'll see that each unit of change for Spectralis is

different than the other two instruments. Spectralis is very close to the Cirrus. A change in the unit of the Spectralis about worth 1.061 -- worth about 1.06  $\mu$  with the Cirrus. The RTVue is a little bit different from both the Spectralis and the Cirrus, with a much wider proportional bias, which is the number that's multiplicative to the device and the formula.

Buchser and Gadi from Joel Schuman's group have recently also published a paper in *IOVS*, looking at RTVue -- excuse me -- Cirrus and RTVue and there's one of the 3D OCTs, I believe. And they found very similar findings. The calibration equations, using very similar -- using a structural equation modeling technique are very different though. And it's sort of interesting that the actual values within the calibration equations don't agree between those two studies, and I'd be interested to hear their thoughts on them.

So, in summary, the retinal nerve fiber layer. The Spectralis had the best inter- and intra-operator repeatability without RTA, and that improved with the scan location. Scans are best performed by the same operator and the same device. Precision is lower with glaucoma patients, worsens with disease severity, improves with eye tracking, and there's greater agreement within Cirrus and Spectralis than there is with RTVue.

So looking at the macula, Garas and colleagues in 2010 again examined -- this is the same study as before. They also looked at retinal nerve fiber layer and looked at the effect of severity on the coefficient of

variability in a repeatability study. The data here is the coefficient of variability. The superior ganglion cell layer scan and the inferior ganglion cell complex scan showing a greater -- a little bit higher reproducibility as you get increasingly severe glaucoma, and some variation between the superior and inferior poles.

This agrees somewhat with the registration data, with a low coefficient of variability and repeatability of 1.9 in normals and 1.4 in glaucoma subjects. Again, this is kind of flipped from what you would expect. And a significant difference, I think, in repeatability between glaucomas and normals, with the normal subjects actually doing worse than the glaucoma subjects in coefficient of variability in the registration data. And I don't know that I fully understand why that is, either.

So the Garcia-Martin study, that also looked at retinal nerve fiber layer, looked at the same patients doing reproducibility and repeatability using a macular scan. Seventy normals, similar to the other study. The coefficient of variability was 0.6, and the reproducibility is not reported as a coefficient of variability but was shown in Bland-Altman plots here.

Hagen and colleagues in 2011 compared reproducibility in normal subjects between two scan patterns within the macula and found a slightly lower coefficient of variability with a denser 512 x 128 scan compared with the 200 x 200 scan.

Lastly, this Pierro study also looked at the repeatability between devices with two scans within the same day. This is the same study that I showed you earlier for the optic nerve and retina scans. And these are the scan patterns for the macula. The Spectralis showed the best reproducibility, similar to the RNFL. RTVue showed the worst repeatability, with Cirrus in between. So the mean agreement, very similar to the other studies, and Cirrus and Spectralis tend to agree. RTVue is a little bit worse.

Looking at total retinal thickness, Wolf-Schnurrbusch's group looked at the Stratus, Spectralis, a Spectral OCT, Cirrus HD-OCT, the Copernicus system, and the RTVue, and very similarly found a lower coefficient of variability with the Spectralis and worse with the RTVue and the Cirrus.

Lastly, Garas -- I think this is my last one -- in 2010 looked at repeatability and reproducibility with the RTVue. A macular Cirrus scan with a GCC similarly found a coefficient of variability of 2.4  $\mu$  and 1.9  $\mu$  superiorly and inferiorly -- excuse me -- percent -- and a reproducibility a little bit higher in both subjects.

One more. Mwanza looked at reproducibility and repeatability in glaucoma patients. This is the same that they looked -- the same population. They looked at retinal nerve fiber layer reproducibility and found a coefficient of variability for the GCL scan of the macula with the Cirrus instrument of 1.8. Within the same population they found a coefficient of

variability in RNFL of .26. So the ganglion sub-complex scan with Cirrus performed better than the RNFL scan with the Cirrus, in terms of reproducibility. This is close to the reproducibility registration data, with a coefficient of variability of .7 for the average thickness of the GCA.

This is just an illustration from their paper showing these tight correlations between all five scans in many of their patients.

So, in summary, precision is lower in glaucoma patients with the macula, just like it is with RNFL and optic nerve measurements, except for the RTVue registration data. The precision worsens with disease severity, just like the other measurements. Precision for retinal nerve fiber layer thickness and macula thickness and ganglion cell thickness is better with Cirrus than the RTVue. Ganglion cell reproducibility seems better than RNFL reproducibility, at least in the Cirrus instrument, and repeatability is better with the inferior -- generally better with the inferior ganglion cell complex scans with the Cirrus instrument.

And that's a lot of data.

(Applause.)

DR. WOLLSTEIN: Thank you.

And the last speaker in this session is Dr. Budenz from the University of North Carolina that will describe for us the implication of the measurement variability of OCT for detecting changes in the different scanning locations.

DR. BUDENZ: I'd like to thank the program organizers for inviting me and the FDA for focusing on our discussion with their questions and definitions. It was very helpful.

I've been assigned the task of reviewing the variability of peripapillary RNFL and optic nerve head measurements and translating this into how we might use this for detecting change clinically.

My department does receive unrestricted research support from Carl Zeiss Meditec, but I have no personal financial disclosures.

We've already heard that glaucoma is a disease characterized by the accelerated death of retinal ganglion cells, and SD-OCT can measure the thickness of the retinal ganglion cell, both the ganglion cell bodies and the retinal nerve fiber layer, accurately and reproducibly. Therefore, we should be able to diagnose glaucoma progression by detecting significant change in the retinal nerve fiber layer beyond the change expected by test/retest variability and normal aging, which is what I'm going to focus on, the test/retest variability.

The next three slides indicate why it's important, in an individual patient, to know the test/retest variability. This is a hypothetical measurement of glaucoma, where it's changed from Exam 1 to Exam 2 by five units. And if this is all of the information you knew, you might consider that this patient has progressed with glaucoma. However, if you know that the variability of the test at each individual exam was 14 units, then that test

change of five units between exams becomes insignificant because it might have been within the variability of the instrument and testing, whereas if the variability is small, let's say four units, five starts to look like change, and we have to suspect that this is real change rather than just the noise of the instrument.

Having low variability is important both for change, but also in diagnosis. The smaller the variability, the less likely we're going to mischaracterize people as having glaucoma on a single examination.

These two scans are from the same young, healthy myope, done at the same session, and you can see on the left a normal scan, everything is in the green, whereas on the right there's a yellow quadrant and clock-hour, and if all you had was the scan on the right, you might assume that this patient had early glaucoma. The signal strength is the same in these two examinations.

And I always encourage clinicians to get at least three OCT scans when doing this sort of analysis so that you can take into account small variations that might tip you into a suspicious area.

This slide is important because it compares the change over time in an instrument that has higher variability and one with lower variability. So on the left, let's say a measurement such as retinal nerve fiber layer thickness plotted on the Y-axis over time, and an instrument with the variability of 8  $\mu$  to 10  $\mu$  from visit to visit, one would have to have a relatively

large change to really detect change because of the variability.

But equally as important, from normal retinal nerve fiber layer thickness of 100 down to the floor, which perhaps is 50  $\mu$ , one would only have five opportunities to detect change during the course of the disease from normal to severe glaucoma, whereas on the right, in an instrument with low variability, let's say 4  $\mu$  to 5  $\mu$ , not only from visit to visit would you be able to detect change with a smaller change because of the low variability, but there would also be more opportunities throughout the course of the disease to say whether a patient is progressing.

I mentioned the floor effect, which we'll come back to when we discuss limitations. And in some of these instruments, the floor measurement is 40  $\mu$  to 50  $\mu$ , so just keep that in mind. And that's why the graphs are drawn that way.

Well, I agree with Dr. Cioffi that we may be getting to the point now that the hardware and software of these instruments is allowing us to detect clinical change, and part of the improvement has been in the technology. And these three areas have been highlighted in a recent publication in *AJO*, and that is, the availability of point-to-point registration with eye tracking on some of the instruments, fixation-independent scan adjustments that can be made, and post-acquisition processing of data, all are superior to what we had with time domain OCT.

And on the bottom is a figure showing on the left perfect

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

registration during a scan, and on the right the effect of saccades. And, of course, from visit to visit, you can have similar variations in the acquisition of data that are overcome by these three improvements in technology.

This is a table that we developed to determine how best to design studies of reproducibility in OCT measurements. First, to actually test reproducibility, as Gene said, we make multiple measurements of the same subjects of interest, let's say glaucoma, on different days in a short amount of time, that time being so short that we don't think the glaucoma has gotten worse, and to ensure that the lower confidence interval of our inter-class correlation coefficient of .8 is no lower than .75, which is the lowest that one would consider good intra-class correlation coefficient.

We design our studies to have three to five repeats, that's three to five separate days, and 40 to 50 subjects. And so this is just derived from standard biostatistics textbooks, Fleiss, in the 1980s.

And the measurements that we use have been reviewed by Dr. Girkin. The intra-class correlation coefficient is what the biostatisticians like because it's not dependent on the mean of the measurements. Coefficient of variation is more intuitive. The closer to zero, the better the reproducibility. But it is affected by mean thickness, potentially, although our studies and others that have been reviewed show that the standard deviation of repeated measures really does not change with the severity of disease. So that means that coefficient of variation might be more applicable or as

applicable as ICC in this particular instance.

And then test/retest standard deviation is used to calculate what I would call the number of what we consider clinically significant change, which is called the upper tolerance limit. And this is a single number in microns beyond which change isn't likely or would occur by chance in fewer than 5% of patients.

And I would disagree with Dr. Girkin. This is the most interesting part of my talk, which is the formula to actually derive the upper tolerance limit. There will not be a test later.

But just to show you how we do that in practice for mean retinal thickness or average retinal thickness with the Cirrus, we start with the standard deviation of the five repeated measures, and we end up with a micron value beyond which is beyond the noise of the machine and one can say with 95% certainty represents change, all of the confounding factors being controlled for.

To review the studies in this light on RNFL, which I've been asked to do, this is a table showing retinal nerve fiber layer thickness reproducibility, and the terminology that Gene used is used here. This is five repeated measurements on five different days, by different operators in a short amount of time, within two months, in 55 glaucoma subjects, which is what we're here discussing. We're not really discussing normals. And just highlighting average retinal nerve fiber layer thickness, the inter-class

correlation coefficient, the closer to one, the better, is .97, coefficient of variation under three. And then the tolerance limit, the upper limit of variability that we expect by chance, is 3.89.

And I've been asked to review similar data with the other two devices.

So with the Spectralis, the overall global retinal nerve fiber layer thickness, which is analogous to average RNFL thickness, the inter-class correlation coefficient is outstanding, .996, with a coefficient of variation less than two. And then for RTVue in glaucomas, shown in the lower right, the ICC is also .97.

I would just point out, and I forgot to do this, but this is a repeatability study, not a reproducibility study. This is a repeatability study of the instrument on the same day and three repeats the same day in 33 glaucoma subjects.

And then this study is 38 glaucoma subjects, also repeatability, which is three repeats on the same day.

These are other studies in the literature that you can review on the main three instruments here in the U.S. There are also some instruments from abroad that aren't available to us, and then the new publication that's in press, comparing all the instruments.

But we can derive some generalizations on RNFL variability from all of these studies, and those are that the coefficient of variation is

quite low for average RNFL, goes up a little bit to 3% to 6% for quadrants or larger sections -- or smaller sections than average RNFL, and then for clock-hours or smaller sections, it's 5% to 10%.

So the principle is, the smaller the piece of the pie, the higher the variability, and that's really due to signal averaging. If you can imagine an individual point, trying to get the reproducibility of that individual one out of 44,000 points, the reproducibility is going to be poor because it's going to be hard to get that actual point. So that's why the average RNFL is more reproducible than the smaller pieces of the pie.

Different regions of the RNFL have different variability, so it's not just a size or averaging issue. But clock-hours and quadrants in the nasal area have less reproducibility than in other areas. And it is true that instruments with eye tracking or vessel registration seem to have lower variability when compared to each other.

I've been asked to discuss the limitations of the studies in this area, and one is that only good quality scans are included in the analysis, which may influence the upper limit of variability. So caution should be exercised in diagnosing glaucoma progression if you're mixing poor and good quality scans.

And I think Dr. Vizzeri has highlighted that poor signal strength gives artificially thin retinal nerve fiber layer. And if we're making decisions on progression based on 5  $\mu$  to 10  $\mu$  changes, we want to make sure we're

using scans with excellent signal strength, 7 or above with the Cirrus, for instance.

And the second point is, that because of the floor effect of measuring retinal nerve fiber layer, we may not be able to use these instruments for RNFL in moderate to severe disease.

These slides were lent to me by Richard Lee. On the left you see normal human retinal nerve fiber layer and ganglion cell layer, and on the right you see a patient with what we would call absolute glaucoma, completely blind from glaucoma, with residual astrocytes and glial tissue that will be measured as part of the segmentation algorithm in OCT. Therefore, because the retinal nerve fiber layer is made up of 30% to 40% glial elements, we will have a floor effect beyond which one cannot measure change.

Secondly, I've been asked to reduce the optic nerve head reproducibility, and I'm highlighting cup-to-disc ratio, which seems to, as Dr. Girkin said, have the best reproducibility. And for cup-to-disc ratio in Cirrus, that's the average of 180 measurements of cup-to-disc ratio around the disc. The ICC is close to 1 and the tolerance limit is .02, as Dr. Girkin said. So one can imagine a strategy whereby a change in cup-to-disc ratio of .03, from .7 to .73, let's say, would be considered a statistically significant change that would occur by chance less than 5% of the time.

Other authors -- this is by Yang et al. -- have looked at optic nerve head reproducibility and found similar excellent cup-to-disc ratio

reproducibility. A third, using RTVue, has already been reviewed. And Spectralis does not have optic nerve head measurements yet, so I did not include that technology.

The FDA questions, one of the FDA questions that was submitted to us was do we need to design studies studying reproducibility with different instruments and operators? And I would submit, yes. We published this study last year. Of 37 subjects, 20 normals and 17 glaucomas, a complicated study design in which all of these subjects were rotated through two different Stratus instruments, two different Cirrus instruments, and were subject to study by two different operators on the same instruments to come up with measurements of inter-operator variability and the contribution of inter-instrument variability.

And what we found was very little -- this is in normal eyes -- very little effect of operator or instrument in either time domain or spectral domain OCT with the exception of, in the glaucoma group, there were some inter-instrument differences in reproducibility. So comparing one Stratus instrument measurements to another Stratus instrument measurements by the same operator, there were statistically significant differences. But this wasn't found for Cirrus OCT or spectral domain in this case, which uses a post-processing way of controlling for registration.

When we look at the variance components in the study, overall, the contribution of operator and instrument, shown in the third column from

the left, is virtually zero. So all of the variability is coming from subject variability, not from differences in operator or instruments, whether you're talking about time domain or spectral domain. And there are two other publications dealing with Cirrus doing these particular studies.

The question was submitted to me regarding physiologic variability. And Dr. Vizzeri mentioned this briefly. But we found no effective large changes in IOP on retinal nerve fiber thickness changes, and the reproducibility on different days suggests very little physiologic variability from things like hydration, diurnal changes, blood pressure, because the ICCs are approaching one.

The one biggest impact of variability is definitely different OCT manufacturers, because of the differences in algorithms for segmentation. And there are multiple studies listed here, showing that you just cannot compare RNFL measurements between manufacturers.

So the implications for detecting change clinically, in conclusion, are that as long as one uses the same manufacturer's instrument that has been properly calibrated and employed, progressive glaucoma can be identified using spectral domain OCT measurements of RNFL thickness and optic disc parameters.

And while longitudinal studies are important to validate this approach and are being performed and published, this method of using reproducibility to assume clinical change has already been used in glaucoma.

The precedents include the Early Manifest Glaucoma Trial, which used the same methodology for validating change, and also the GPA analysis of the Humphrey-Zeiss field analyzer used the same study design of five repeat measurements over two months in glaucoma subjects, to document what the noise is in the instrument to determine what changes.

Thank you.

(Applause.)

DR. GREENBERG: Well, thanks very much, Don.

At this point, what I'd like to do, Don, and all the other speakers, is invite you to please join us here for a panel discussion. We've had some fascinating talks. I want to congratulate all our speakers for tackling very complicated topics. I'd also like to invite a few others to join us on the podium: Bob Weinreb, Doug Rhee, as well as some representatives from the FDA, Dan Hammer, Kristen Meier. This was a very complicated topic to discuss and to digest, and I congratulate our speakers.

There are microphones throughout the room here, for those in the audience that would like to raise questions. We do have participants who are listening to us on webcast. It's possible that questions may arise. If they do, I'll ask those to be please passed over to us so that we can please include those that are on webcast, that relate to this particular session only.

We did speak about three specific topics in this session. We talked about sources of artifact that can compromise quality. We talked

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

about variability and we discussed, as well, some issues that pertain to the regulatory science and the regulatory approval process as it relates to these devices.

The first question that I have is actually for Dr. Hilmantel, and it relates to some of the regulatory issues.

You brought up the definition of a predicate device as a legally marketed comparator, for example. And I was wondering if you could perhaps clarify for those of us that are less familiar with the process of an approval from the perspective of the FDA, what would be a specific example of a predicate device as a comparator for a new spectral domain OCT instrument? What kind of device are we speaking of?

DR. HILMANTEL: Okay. Well, really it's defined by the whole 510(k) process. So a predicate device is chosen by the company, by the sponsor, and it can be any product that's on the market that measures similar things. So it doesn't have to be a specific spectral domain OCT. It can be any type of OCT, really.

DR. GREENBERG: There are many people here who are representatives of industry. There are many clinicians and scientists, all of whom have participated in the approval process, for example, of normative data.

And I'm wondering, Gene, if you could provide some additional guidance from the FDA's perspective. Is there a recommended measure of

reproducibility? And can you perhaps provide guidance on what the minimum levels of agreement are from the perspective of the Agency?

DR. HILMANTEL: We don't have specific levels that are required, okay; it's all a matter of comparison. The sponsor provides the clinical data from the new device and compares it to the collected data from the old device. So they'll compare the reproducibility of the new device. Is it similar to that of the old device? Are the measurement values similar to the old device? And we make an assessment based on comparing the two. That's defined by the premarket notification process. So we don't have a specific bar for that.

DR. MEIER: This is Kristen Meier.

If I could just add to that. I think we have some questions that we're actually asking the panel here about what kinds of differences are important. That's actually why we're here, to actually ask here. We do actually do comparisons of precision. So, again, in picking a predicate, I think it is helpful to pick a similar technology because we will be comparing -- or the manufacturer would be comparing the precision of their device with that of the predicate. So you would probably want similar technologies because they would probably have similar precision.

Again, I think that's the question that we are actually bringing to this group of experts here, is what kind of differences are important?

DR. GREENBERG: Yeah, you bring up a good point, and it

probably is a point that many of us don't have necessarily any real answers for. You're raising a question of what might be clinically relevant. We also don't know what might be relevant from a regulatory perspective, which is why I raised the question. And we saw the data that's been published to help clarify what we can expect from some of these devices.

Doug Anderson, did you have a comment on this topic?

DR. ANDERSON: Yeah. Before you get away from predicate, I thought I would try to clarify a little bit about that.

Let's imagine, for instance, that we have an OCT device and the OCT device itself is already on the market, but it develops a new measurement. And let me say, for example, that retinal ganglion cell volume has something that hasn't yet been measured by this instrument and someone goes ahead and says okay, I'm going to start measuring that, and now here are my data on what we can do. We can make these measurements, and here's the reproducibility of it, and here is the range of normal, and here's the range over the course of disease. But there's no predicate device that has made that measurement.

How are you going to judge that? What becomes the predicate for that?

DR. HILMANTEL: Yeah, that's a good question. That actually comes up quite often, as you can imagine.

So, basically, the sponsor just has to provide some sort of data

demonstrating the validity of what they're measuring. If they're measuring a new parameter, they can give us data, for example, from, like I was talking about in my talk, segmentation studies, things like that, things where certain parameters may be measured manually from the images as designated by experts. The experts may make measurements on images and then, whatever the device is measuring, it's compared to that.

DR. ANDERSON: Is there an element at all in this, not of just being able to -- let's say the predicate could be something that's not on the market, namely a manual segmentation, you said. That's not something on the market, and so it doesn't quite qualify as being a predicate by definition.

But is there are a place there to say that this is helpful in the course of making a diagnosis or in following the course of the disease? That is to say, let's say your only evidence is that in people with early disease, you have this range of numbers, and people with moderate disease this range of numbers, and in severe disease this range of numbers. And therefore this new measurement that's never been made before by anybody, by any instrumentation, is useful in telling the course of the disease. How do you deal with that? Because there really isn't a predicate, as I see it.

DR. HILMANTEL: Malvina wants to comment.

DR. GREENBERG: Can I also remind all the speakers to please provide their disclosure? I think, Malvina, we all know your disclosure.

DR. EYDELMAN: So I just wanted to make a couple of

comments just for purposes of clarification.

In light of your statement about five minutes ago -- that's when I popped up. I just wanted to make it clear that we're here for the main purpose of getting your input. The regulations don't have -- I mean, if you open the C.F.R., it doesn't say the precision for OCT devices has to be X, Y, and Z. That's where we come in. So what we're asking for is your clinical input so that we can utilize it to make our regulatory decisions.

Did that clarify it? I thought there was a little bit of confusion.

So what Gene was alluding to is no, we do not have a guidance that currently delineates numbers for precision.

DR. GREENBERG: Well, is there a threshold above which you would consider a device to be so poorly repeatable that it would not necessarily be even considered? Is there a specific range within which you consider devices to be acceptably reproducible?

DR. EYDELMAN: So as I was trying to allude to in my presentation, each device is being evaluated in comparison to the predicate. So the sponsor makes the case why the two devices are substantially equivalent.

The main purpose of today's meeting is to exactly get the input of what you're asking us to give answers for right now. So we're trying to be consistent and take your input and come up with the criteria that will be clearly communicated then to the community.

DR. GREENBERG: Chris and Don, Kristen and Malvina would like some guidance on what is clinically considered to be acceptable reproducibility and repeatability.

Chris and then Don. Or Don and then Chris.

DR. BUDENZ: You know, we have what the statisticians have defined as "good" and "excellent" and "fair" reproducibility by ICC. Those are the only definitions that I'm aware of, and I would hesitate to, you know, apply those clinically without some sort of consensus. But there is the ICCs; .75 and above are good, and I guess it's .85 or .9 and above are excellent. But those are relatively arbitrary statistical definitions of what's excellent.

DR. GIRKIN: I guess bringing it forward into the clinical realm, it really depends on what the tolerance limits are for change, like how many microns exceeds that 95% confidence interval. I think it's 1.5 in your study or something like that for RNFL. And the more precise it is, the more efficient you can detect change, and the less precision it is, the less efficient.

So if we think about it that way, we can think about, you know, what sort of level of precision, in terms of microns, in terms of the range of microns, is important to the clinician, and then you can convert that back to a level of coefficient of variability.

So I mean, I don't know what that level is, but is it important in a global RNFL? You can detect a global RNFL change, a real global RNFL change at 1  $\mu$ , 2  $\mu$ , 3  $\mu$ . Then that brings the question --

DR. GREENBERG: To narrow the discussion, we know from your presentations that, as we use and discuss more broad parameters, the repeatability is better than, comparatively, if we focus on quadrants and sectors.

So if I propose we're talking about the average retinal nerve fiber layer thickness, for example. Can I ask Joel, for example, to comment and then Bob Weinreb?

For that particular parameter, what's a clinically acceptable level of reproducibility?

DR. SCHUMAN: I think that, you know, the smaller variability is always better. But there was a time when the variability was -- the standard deviation was 10  $\mu$  and we worked with what we had.

And so I'd like to just make sure that we all understand that we're talking about two different things. One is the FDA regulatory function, and then the other is what we would like to have as clinicians and scientists. And the latter really is where we're talking about the 1  $\mu$ , 2  $\mu$ , whatever. In the regulatory, I don't know if we want to suggest that the standards need to be that strict.

DR. GREENBERG: Bob Weinreb, any additional thoughts on this topic?

DR. WEINREB: I think Joel's comments are right on. I think, as clinicians and as scientists, we need the smallest possible number. I think we

need to be careful not to be too restrictive with any regulatory types of policies.

DR. GREENBERG: I found it interesting, in the discussions, that the coefficient of variability was the highest for the measurement of cup/disc ratio, an assessment that we feel, as clinicians, is probably the most poorly reproducible.

Is there any feeling about whether or not that implies a sense that that might be a more useful parameter, as a clinician, because of its repeatability compared, for example, to the retinal nerve fiber layer, or less useful because of the nature of that parameter?

DR. HOOD: This is where you have to take into consideration the difference between precision and accuracy.

DR. GREENBERG: So Don, in your presentation you spent a lot of time talking about segmentation as it relates to quality and an instrument-dependent factor that affects the output.

Is there any data about the reproducibility of segmentation algorithms within a given instrument, for example?

DR. HOOD: Well, it depends on what you mean. If the algorithm is given the same image the second time, it damn well better do the exact same thing.

DR. GREENBERG: You showed an example where there was some artifact produced by blood vessels.

DR. HOOD: Yeah. Well, it's going to do the same thing the second time to the same image. Now, the question is what happens if you scan again? But then, now you're talking about reproducibility of the hardware at that point and not reproducibility of the algorithm.

Does that make any sense?

DR. GREENBERG: Yeah. But is there data that specifically looked at the ability to segment the same retinal locations, as we have data for reproducibility of specific parameters?

DR. HOOD: Well, that's going to be built in, in a sense, to the studies we just heard about.

DR. GREENBERG: Yeah.

DR. HOOD: Because it's a combination, you know.

DR. GREENBERG: Yeah.

DR. HOOD: It's a combination of both the reproducibility of what the hardware is doing and then how --

DR. GREENBERG: It would be.

DR. HOOD: -- the scan may differ if you change contrast or if one layer is a little fuzzy or something like that. So it confounds both of those factors. That's why I'm saying, you know, the big difference to me -- at least I learned a big lesson from just comparing the time domain with the frequency domain, and I'm assuming that's true with the other machines, too. The big factor here is the difference in the algorithms.

DR. GREENBERG: Joel, you know, in your presentation you showed the evolution of OCT over the last two decades, beginning with time domain OCT.

Can you comment on how that evolution has impacted the reproducibility of the assessments?

DR. SCHUMAN: I think that it's clear that we can make more precise measurements. I can't speak to whether or not they're more accurate, but they're certainly more precise.

And so if you looked at the reproducibility of the prototype unit, if you say what's the standard deviation of measurements for the mean nerve fiber layer thickness, it was about 10  $\mu$ . If you look at Stratus, it was time domain, commercially available, best available, it was about 2.5  $\mu$ . And now with most of the machines, we're down at about 1.5  $\mu$  standard deviation of measurements.

So we're seeing more reproducible measurements as the technology improves, which is what you would expect, especially if what you're doing is being able to register the images. So you're pretty much measuring in almost exactly the same place every time.

And what that does, I don't think it helps with our discriminating ability, which is really what we're talking about today. But I think that it will be helpful in a longitudinal analysis of our data, so we can detect smaller changes over time.

DR. HOOD: Do you think we should try to answer the FDA questions?

DR. GREENBERG: Yeah. Well, we also have questions from the audience.

Is there another question from the FDA that has not yet specifically been addressed in any of the discussions, that would like to be further discussed in this panel?

DR. HILMANTEL: Well, I'd like to make a couple comments here. First, the reproducibility issue also plays in heavily to the normative database aspects that we'll be discussing later. So a device may have what looks like pretty good reproducibility, but sometimes a relatively small change will cause a large percentile change compared to the normative database. So that's a big issue, too.

But as far as our questions, we're just looking for some guidance as to also what is appropriate to include in the labeling for users, whether -- and I gave a couple of examples of the types of things we put in labeling -- whether there's anything else that the community would like to see in the labeling.

DR. MEIER: And Gene, if I could just add to that.

Also, I just want to be sure I'm clear on 3.3.a, about the sources of variability. I heard that, at least in some of the studies done, there wasn't a lot of device operator variability. I wasn't sure if those were highly trained

operators or really sort of a broad spectrum of people with differing amounts of experience. And also about the day-to-day variability, I wasn't sure if I heard conflicting information or not on that. But if folks could just reiterate their thoughts on 3.3.a, that would be helpful for me.

DR. VIZZERI: So I can comment on this particular question.

I think we were in agreement, Dr. Budenz and I, with regard to the first question, operator variability. You raise an excellent point. For example, with Stratus OCT, what I've experienced in the clinic is that with non-experienced operators you can have higher variability. So having a more operator-independent technology is certainly the other thing, in my opinion. And a lot of the studies come from groups that had very experienced and skilled operators. But time domain OCTs, at least based on just the way the technology functions, are more operator dependent.

DR. GREENBERG: As a point of clarification, we've been using phrases that are rather subjective, like better or worse, and it's useful to point out that something might be better statistically and not necessarily better clinically. For example, we saw data that tracking considerably improves the repeatability of the measurements considerably, 1%, for example, COV as opposed to maybe 2% or 3% without tracking. And while that's statistically better, it's to date unclear if that's clinically better, although we would expect that it would translate to something more meaningful clinically.

With regard, Kristen, to your question about the experience of the operator, there is data in the literature over the last 15 or more years, particularly in the early introduction of time domain OCT, with regard to inexperienced operators, where the parameters had coefficients of variability on the order of 3% or 5%, which are not terribly different than some of the data that we've heard presented today with spectral domain OCT.

I don't want to ignore the audience. We have several people at the microphone. If you could please introduce yourself and then present your question. And please, again, remember to include your disclosures where relevant.

DR. BUCKLAND: I'm Eric Buckland. My disclosure is that I'm from Bioptigen.

So a couple of things occur to me. One is the discussion on precision and variability. If we start with precision, one thing, it's not just the algorithm, right? There are calibrations both in the axial dimension and in the lateral dimension, and those calibrations are completely proprietary between companies. So there is no way to compare how the axial eye length is used to translate from the angle, which is what is scanned in the eye, to the lateral dimension.

So some disclosure on how companies make that translation would help to make sure that we have lateral dimension equivalence, which means that the circumpapillary scan would be in the same place.

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

Similarly, in the axial dimension, there are differences just in calibration that can be just a systematic bias. And the companies that are very close together, that you indicated, clearly have done a lot of work to make sure that they're the same, and there's another that's different.

There are ways, I think, the objective of having a phantom that is -- and I don't think it has to be as complex as a retinal look-alike, but providing just some known optical distance to validate calibration is a very easy thing to do, and that could be disclosed, as well.

But two other comments or questions. It would be very nice if -- there is a lack of guidance to companies. We don't want to have a subjective process as we bring new devices and measurements to the market and to the FDA. And in the lack of precise guidance, there are at least a couple of things that could be offered. For example, define the layers. Define the layers so we know what we want to measure between. So as we know, to get to retinal thickness, there's a big difference between which of the outer retinal layers we're using. Those are choices. So if we all settle on the same place, we will have a much, much lower bias.

And then the last comment. The process of developing normative data and validating measurements against predicates is exceptionally complex and costly. And as we know, the particular protocols that are used by the different companies, those are also not public, which means that every company has to kind of reinvent the right way to develop

those methodologies.

So bringing some transparency to those sorts of items, which doesn't require, I think, a full guidance document on everything OCT, would go a long way to making sure we can get more measurements, more tools out faster, higher precision, lower cost, and to cover more clinical avenues for you.

DR. GREENBERG: Thank you very much for your comments.

We are running a bit late here. We have time for two very brief questions from the audience.

DR. DAVIS: Sure. I'm Pepe Davis. I'm affiliated with Carl Zeiss.

My question was really following up with regards to the recommendations for reproducibility and repeatability limits for comparing to predicate devices.

Could the panel maybe recommend that as long as the data is disclosed or the results are disclosed, as we are doing, that that is sufficient since we can't come to a particular limit with regards to how to do the comparison? Would that be sufficient?

DR. GREENBERG: I'm sorry, to repeat your question, are you asking for a consensus among the members of the panel with what the threshold tolerance would be for variability?

DR. DAVIS: I'm suggesting that maybe there isn't a threshold. Maybe as long as it's disclosed, what the variability and the repeatability is,

that maybe that's sufficient from an equivalence standpoint, rather than trying to make a particular limit since we couldn't seem to get to a particular number or value.

DR. GREENBERG: Any comments from the panel?

DR. HOOD: It almost sounds like it makes sense, with the argument against disclosing that information.

DR. SCHUMAN: Well, doesn't the FDA regulate the quality of the instrument that's being put out there?

And so if your instrument had a very large variability, obviously that would be undesirable, and so just disclosing what the measures were probably wouldn't be adequate. But it's very hard for us, as a group of clinicians and scientists, to come up with what's the right number that the FDA should choose.

DR. HOOD: I think he was asking, at the very least, shouldn't that be disclosed?

DR. SCHUMAN: I thought that was -- then I heard you wrong, because what I heard was, would it be enough to just disclose it?

DR. DAVIS: So my suggestion would be that if there is no particular limit, then if it's disclosed and the clinicians can make up their minds as to how to use the device, given the limitations and the data presented -- I mean, depending on what diagnosis you're making, maybe a certain amount of variability is tolerable within the clinical decisions versus

trying to make a different decision. Then you would need less variability. That would be the suggestion.

DR. MEIER: If I can just interject. This is Kris Meier.

That information is disclosed. And, again, I think that was some of our questions. How should that be presented? You know, there are lots of ways to disclose information, okay, and what's a meaningful way to disclose that information is actually what our questions were.

DR. GREENBERG: We do have time for one very brief question from the audience.

DR. MORAES: Gustavo de Moraes. No disclosures.

My question is actually more of a suggestion. We've been talking about variability in a normative database, which refers mostly to cross-sectional analysis, if you want to diagnose glaucoma or not. And I didn't see anything related with longitudinal changes, how the variability changes according to disease severity. We have a lot of knowledge from perimeters regarding that. And I don't think, even though there's a lot in the literature, and that Don Hood has published a lot on that, I don't think any of the companies now address this issue of short-term variability depending on the severity of the disease over time.

DR. GREENBERG: Chris, you showed some data on glaucoma variability being greater than normals. Is there data on the degree of severity?

DR. GIRKIN: Yeah, there are two papers looking at, I think, RNFL and retinal thickness that I showed. I need to refresh my memory on the slides. They just really broadly broke up groups into normal mean defect from 6-12 and then > 15. I'm not sure why that definition of severity was used. And it did increase. And I can't remember off the top of my head, but I think it was on the order of, like, 1-4 in a coefficient of variability across the range, or something in that magnitude. So it does affect it, but you know --

DR. HOOD: I'm doing it from memory, but we published data on Stratus, and my memory, I mean, we certainly compared to fields, which has this big effect reasonably constant or relatively constant. I'd have to go back and look and see how constant.

DR. GIRKIN: We did it with HRT, GDx, and Stratus, as well, and found very similar layer change thus far.

DR. GREENBERG: Right. Well, I want to thank our excellent speakers, and I want to thank all of our panelists for really providing their expertise to us and to the FDA.

That concludes this morning's session. We're going to break for lunch until 12:30 and reconvene exactly at 12:30 for the remainder of the afternoon session.

Thanks very much.

(Whereupon, at 11:30 p.m. a lunch recess was taken.)

AFTERNOON SESSION

(12:30 p.m.)

DR. KAHOOK: We're charged with covering construction, reliability, and usability of the devices being discussed.

My name is Malik Kahook. I'm Professor of Ophthalmology at the University of Colorado, and I have no relevant disclosures for this session.

My co-moderator is Professor Shan Lin from the University of San Francisco, and he'll be introducing each one of the speakers for the session.

We've assembled an expert panel to cover each section, and we will then open up the mikes for 30 minutes of discussion, similar to the last session. I want to remind each of the speakers to stay on time so that we can cover all of the information that we're tasked with covering. And without any further delay, we'll get started with our first talk.

DR. LIN: I have no relevant financial disclosures.

Our first speaker will be Rob Fechtner from the university in New Jersey, and he will speak on the "Review of Normative Database Construction in Available OCT Models Highlighting Differences."

DR. FECHTNER: Thank you, Shan.

And I very much appreciate all the people who put so much effort into getting this meeting together. I know we have an audience out there on the web, so Hi, Mom. And let me get started.

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

The normative database construction in available OCT models was my topic, and as I started speaking to the experts involved in their construction, it struck me that we have a heritage and a language we've been using surrounding images and imaging technologies that sometimes do us a disservice as we're thinking about the scientific aspects of it.

I have been around the development of a number of these instruments. Things are different now. We have greater experience, we have different approaches, and yet, as you look at what we do with our imaging, it is imprinted with the entire history of the computer-assisted imaging technologies.

I have no financial interests relevant to any of the content I'm going to present.

So what's in a word? Well, I'd say one of the first things we've been saying today is a normative database, and that may be the wrong word, because it gives us the wrong concept. What we have in a database is a reference. It serves as a reference point as we compare an individual. If all we did was image an individual, we'd really kind of be in a vacuum at interpretation.

So I'm going to talk about reference databases. And once we think about them as reference databases, our expectations change. We see this in laboratory medicine. I no longer get the normal limits; I get the reference limits.

So if we're talking about a reference database, what does it do?

It serves as a reference for a new examination. It can determinate percentile rank for one of the calculated or measured parameters, much as a standardized test score would, which we've all taken.

But the reference database is a tool, it's only a tool, and its results must be interpreted in the clinical context. I think that was a point brought out in the last panel discussion. The database will be a reference. What is done with that reference is clinical decision making.

So what the database can answer for us is in what percentile does the measurement fall, compared with the reference population? I don't know anything that it can do beyond that. It's a rank.

What can it not do? It does not answer the question, Does one have disease or not? We know that there is overlap any time we try to categorize.

We've been done, perhaps, a disservice in our concept of looking at our reference databases and thinking normal, borderline, abnormal, in anything other than a statistical sense. You can't make the clinical decision, but you certainly have a statistical indicator there. And if it's a properly designed database, then you're more likely to have disease if you're way out at the tail of distribution.

So what is a reference database? Well, we have to ask the question, who is in the database? And who made the decision to include or

exclude? Is it a very narrowly defined database or is it a broad population-based database? They can serve different purposes, but they may not have such very different effects, in the end, for clinical use. And if we do exclude or include, based on what criteria are they valid?

In order to try to get this information, there really is no single source. I went to the FDA. There is information on the web. By going and looking at those premarket notifications, you can see the letters. We reviewed all of those. We looked at the published literature.

For Optovue, I also had some personal communication. With Cirrus, there was published literature at the FDA website and personal communication; and with Heidelberg, the FDA information and personal communication.

So here's what I was able to discern about the approved databases that are available. There are different numbers of study sites, and there's different geographic distribution. For the RTVue, they had various sites around the world. Not all of the sites contributed data to the database that's currently on the platform. Cirrus had six U.S.A. and one China. The Spectralis database was generated from a site in Germany.

There are different numbers in the database, and there are different subsets of those numbers in any of the particular parameters, depending on the quality of the examinations available, at least on some of the platforms. So we don't have precisely those numbers for every single

parameter.

Various ethnicities were included. The age ranges were very similar, from just about the age at which you can give consent to be in a study in the U.S., of 18, to -- not the upper limits of glaucoma, but to elderly -- in the 80 or so range. And then significant limits are included in them, and there are different significance limits based on the different databases. You can read from the slide what those significance limits are.

Exclusions were part of the process of these reference database collections. If you look, I categorized them as ocular history. So we're now narrowing down. Some questions I had were about the particular conditions. There were reasons these were used as exclusions.

But I will ask, later, a question: What is the consequence of exclusion, their benefits and their penalties?

Medical history was used as an exclusion for constructing the reference database. You can see what the conditions that were listed in the information available to me were. And at least for one of the databases, they excluded a family history of glaucoma in a first-degree relative. There were also medication exclusions in one database. As we know, some medications can affect the retina and the macula.

Refractive error was used to create boundaries. Certainly we know, as clinicians, that the optic nerve can look quite different in a high myope, and there are magnification issues.

Intraocular pressure was used as a parameter for all of the reference databases, with a fairly typical cutoff level of  $< 22$  or  $\leq 21$ . Really the same cutoffs.

Visual acuity was used as a parameter. The subjects in the reference database all had really quite good visual acuity.

It is a well-discussed issue in our literature for imaging, do you or don't you use optic nerve appearance as an exclusion when you're looking at parameters surrounding the optic nerve?

And it was approached a little differently on each of the reference databases. For the RTVue, there was no exclusion on the optic nerve. For Cirrus, no disc hemorrhages and no retinal nerve fiber layer defects. And for Spectralis, a normal appearance of the optic disc by two examiners.

Visual fields were used as a parameter. They were graded slightly differently. But for each of the reference databases, a normal visual field by the specified parameters was required for inclusion.

So we end up with a reference database, which in the approved versions are narrowed by the inclusion and exclusion criteria. And then we have a printout, which we've been talking about today, at some points as being normal, borderline, or abnormal.

I think the visual display of the information we have has become very familiar to clinicians. We know what the nerve fiber layer

representation looks like. We know what the TSNIT graph looks like. We know what the optic nerve imaging looks like. And we are very visual as clinicians. I think Jack Cioffi made the point that we have excellent pattern recognition, and we use that.

We've also been given color to inform us about that data from instruments, which are essentially color blind. And as I selected my outfit for today, I'm gray scale.

(Laughter.)

DR. FECHTNER: We have learned how to read gray scale. We are very familiar with gray scale, and we're also very comfortable with the idea that the gray scale is giving us statistical information.

We've moved into a world of full color in our representation of the statistical data for our imaging devices. It was intuitive. It was possible. It was not possible when visual field machines first came out.

So if we will re-conceptualize how we've coded our color perception into statistical, then we're dealing with a very straightforward problem of reference, such that perhaps our greens are greater than or equal to 5%, our yellow might be less than 5%, and our red might be less than 1%. It's an exact parallel to how we view our functional testing.

The structural testing dataset is much more robust. We don't look at the raw data. We only look at the processed data. And it can be represented quite elegantly, but it's a reference point.

By taking a very defined population, we have to ask the question of generalizability. All the patients whom I see do not have acuity of 20/30 or better and do not have normal visual field tests. They have medical conditions. So a reference database which has excluded a population that matches my patients becomes a less useful tool. It's not generalizable.

I took the example of the Cirrus database where 26% were excluded for unreliable visual fields. In fact, one might argue that structural imaging is most needed in patients who have unreliable visual fields where we're not getting the functional data.

Have we sufficiently reexamined the assumptions underlying this approach of our very defined populations, or trying to populate the reference database only with normals? We went ahead and did that. We've done that over our heritage of imaging without really revisiting this question.

So who should be included? Can or should optic nerve abnormalities be an exclusion? Does that bias the reference?

Should medical conditions be an exclusion? Does that limit the utility and usefulness of these devices when our clinical population is aging and has medical conditions?

Is a visual field required? If you have to be able to produce a normal visual field, then how do I apply this information to a patient population who can't reliably use visual fields?

So the more exclusions we apply, the greater the chance we've

created a super-normal reference population. Is this a desirable outcome? I think it's a question that has to be asked as we move forward and technologies evolve.

Why not the inclusive? It would predictably improve the specificity. If we actually have some eyes in there which, for example, might have glaucoma, then maybe your nerve is going to have to be a little bit worse to get out to that one percentile tail, which we will color red.

What if we included everyone? It only could be done if you have an unbiased population-based cohort. Some with disease would be included, but the reference database would still resemble the general population if it's large enough in a disease with low prevalence. And then it might have more applicability when we're using the diagnostic tool against the general population. It's a question which really hasn't been revisited surrounding our reference databases.

In the end, we end up with a percentile. We don't end up with normal, borderline, or abnormal. A reference database serves as a reference for new examinations. It can determine percentile ranks for the measures, and it's a tool that must be interpreted in the clinical context.

The topics we're discussing today, I think, are of great importance to both the regulatory world and to the clinical world as we're developing these tools going forward. There certainly is an aspect of clinical education, which is the responsibility of the clinical educators. But the

science and development remains with those who are doing the building of these tools and the building of the statistical tools to allow us to understand them. And I believe the expertise resides in the FDA and in the AGS, and it will fall to us to provide guidance for acceptable reference database construction. The dialogue is open.

But I believe some basic questions must be addressed, and we shouldn't assume that the way we're doing it now is optimal in terms of efficiency or usefulness.

We could have a model protocol for basic reference databases. And I often think of my kitchen, where I have my knife block. It started out with a paring knife and a chef's knife. I have all sorts of specialty knives out there now, but I started out with some basic tools which let me do an awful lot of cooking. We need the basic tools so we can approach our patients with a reference database which is meaningful and useful. And I believe the world is going to be very interested in the dialogue as they look to the U.S.

There will be a World Glaucoma Congress, and this is a topic that's going to be discussed further at the World Glaucoma Congress, and I hope some of the scientists here will submit abstracts and continue this dialogue going forward.

Thank you.

(Applause.)

DR. SHAN: Thanks, Rob.

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

Our next speaker is Kouros Nouri-Mahdavi from UCLA, who will speak on "Statistical Approaches for Determining Normal Limits in a Database."

DR. NOURI-MAHDAVI: Thank you. I would like to thank the FDA and AGS and the organizing committee for the invitation. It's a great pleasure to be here today.

And my task this afternoon is to continue Dr. Fechtner's discussion with regard to statistical approaches for determining normal limits in the OCT databases.

I have no financial interests with regard to any OCT devices.

And I would like to thank our consultant-statistician at the David Geffen School of Medicine at UCLA for statistical advice and discussion.

Now, I will touch on a few things about detection of disease, which is the main goal of making a normative or a dataset and reference databases. I'll talk about the definition of normal, choice of the sample size, composition of the normative database, and a little bit about covariates as a segue to Dr. Zangwill's talk.

Now, Dr. Fechtner very beautifully described the problems associated with the definition of the non-diseased, as the statisticians like to call the reference group.

And just to sum up, I would say that probably most people in this audience would agree that a normal eye exam, and particularly normal

intraocular pressure and normal achromatic visual field, is probably the best way to go for defining the normative database.

But the main issue actually lies in whether -- or the main question, actually, or controversy lies in whether we should include normal discs and have a super-normal database, or we shouldn't include and have a lower sensitivity.

And I would argue for the second, for the latter, because having a super-normal database would lead to a lower specificity, which is detrimental both in the clinic and in the screening setting, while having a lower sensitivity, which is not as much as one would think, because the prevalence of pre-perimetric glaucoma is actually not that high in the general population, especially in the age range of 40 to 70 years of age, where most of the databases are concentrated on. And given the fact that most of these databases actually are using Gaussian assumptions, it actually is not too detrimental, actually, to the cutoff levels that are commonly used.

So I just mentioned that most of the databases are actually using Gaussian distribution, and it's probably right. Most of the biological parameters in humans actually do follow a Gaussian distribution in one way or another in the original scale or in the transformed scale. But from a purely statistical point of view, nonparametric estimation methods actually are more robust, but do need higher sample sizes, which might be a big issue for creation of these normative databases.

The green curve here actually is showing a hypothetical structural outcome, which is following a Gaussian distribution. As you can see, we really don't care about the top part of the Gaussian distribution. We're actually interested in the bottom part. And the cutoff points are obviously using one-tailed statistics. Going down 1.645 standard deviation from the mean would give us the 5% cutoff level, and 2.326 standard deviation unit would provide us with a 1% cutoff level.

And despite the fact that we tend to think that the statistical analyses are pretty accurate, we should remember that these are all soft cutoff levels. And I will return to this in a minute.

Now, in a customary way, convenience sampling has been the way to go for choosing the sample used for creation of normative or reference databases and obviously because, as it says, it's convenient. It's very hard to run surveys and population sampling for doing these, creating these normative databases. And there's also an issue of generalizability, and therefore we're stuck with the convenient sampling.

Now, the result of this is that we're stuck with two problems. One of them is the selection bias. And selection bias, in and of itself, is actually not a huge problem when we're looking at correlations or predictors of structural outcomes. We know that things like age and other things, axial length and other things, do affect the structural outcomes. That's not a huge problem with these correlations. If there is a selection bias, it's more a

problem with defining the prediction integrals, like the 5% and 1% cutoff levels I just mentioned.

The other question that I would like to answer is one of the questions that actually has come up in the list of questions provided by the FDA, is that the distribution of the predictive factor of the covariate needs to be pretty similar to the ratios in the population, if it warrants, and if we're dealing for prediction limits. Again, it's not an essential part for correlations.

Now, why do we need a big sample size or a fairly big sample size for these reference databases? The problem is that the effect of a small sample size is that the precision for the 5%, 1%, or another cutoff level we would like to have actually decreases significantly if we have a small sample size. The question is does it matter, really, from a clinical point of view?

This is a table that shows actually the confidence interval here for the fifth percentile. And as you can see, with the smaller number of sample, with smaller sample sizes, you have the very wide confidence interval, 95% confidence interval, for the cutoff levels, whatever it is. But somewhere around 240 to 480 eyes, we get to a certain limit where it becomes very satisfactory from a clinical and statistical point of view. The 95% confidence interval here is about a quarter of standard deviation, which seems to be where we really want to be.

Now again, as Dr. Fechtner alluded to, these statistical cutoff points are as they said, they are statistical, they're artificial. Per se, they do

not have much clinical meaning. And I would like to suggest that maybe we should be thinking about a functional definition of abnormality.

And I'm referring here to a variance study by Wollstein and colleagues from Pittsburgh, where they looked at the tipping point. That's where really a change in the RNFL thickness or another parameter actually is leading and resulting in a change in function. And this has not been looked at in detail and needs to be further studied.

Now, one thing we're doing when we're defining all of these cutoff points for average global sectors, quadrants, and everything, we're doing some of multiple hypothesis testing, and therefore the real p-value is actually -- if you look at all of these numbers, it's not really 5%. The real value is way higher than 5% and it's a function of the correlation between these sectors.

And there are statistical ways to deal with it which need to be explored, such as defining prediction spaces rather than prediction lines or intervals, prediction spaces for all of these sectors being abnormal at the same time. And it's more important when there's only a couple of these sectors and the significance is at fairly high levels, say at 5%, 7%, or 10%.

Once we have a very thin RNFL, like this sector here, which shows significant abnormalities, obviously that is abnormal from a statistical and clinical point of view.

So as I mentioned, there are known covariates that affect our

measurements in the reference database, and obviously these reference databases have to be corrected for these potential covariates. This is commonly done by either multivariate adjustment or stratification. And the main difference here is that with the stratification, you have the power to look at particular substrata in which, actually, you don't have enough in the general population.

Let's say that you're interested in the standard deviation and mean of some structural parameter in the general population, but that particular race, for example, is not common in the general population. And I just said that we need to follow distribution for defining the reference database. By stratification, you can actually plan ahead of time to have the higher sample size for that particular substratum and have a higher precision for definition of the reference intervals.

Now, I kind of like to divide the covariates into two groups. There are some that actually affect the distribution of the RNFL or whatever the outcome of interest is. But they're not really associated with worse outcomes such as gender or race or disc size or signal strength. They're not typically associated with a worse outcome.

But other parameters of covariates, such as age or axial length, when taken to the extreme, like here, when you're looking at eyes with high axial length or very old people, do we have to have people of the same age or the same axials in the normative database? Or what we're really looking for,

or what they're really looking for, is to know whether that particular measurement in that particular patient, even if they're very old or have a very big eye, is abnormal compared to the average population, who might be younger or who might have a smaller eye. And that's something we could discuss later during the panel discussion.

So to sum up, the role of covariates actually is to hopefully and potentially improve the precision of prediction intervals.

As Dr. Fechtner beautifully described, I would argue for having a sliding scale significance level, not having particular cutoff points. And there's nothing magical about these cutoff points. And it would be very valuable to have not only sliding scale significance levels, like knowing what is the exact rank of a particular sector or global outcome; also it would be very interesting to have the likelihood ratios for those particular findings.

And I also would argue for use of mathematical modeling, in addition, and on top of the statistical modeling, because there are a lot of things, anatomical factors and parameters in the fundus, that are not easily correctable or adjustable by using regular statistical means.

And just to sum up again, all of these statistical issues and ways to correct and create the best normative database we can are valid only, and if only, the non-statistical issues we discussed in the morning are taken care of. And we're already under way to address those. But these come first, and then comes the statistical improvement of the normative databases.

Thank you very much.

(Applause.)

DR. SHAN: Kouros, thank you.

The next speaker is Linda Zangwill from UCSD, and she'll be talking to us about "Stratification of Normative Data."

DR. ZANGWILL: Good afternoon.

I want to acknowledge my financial disclosures, and that we receive research instruments from several of the spectral domain OCT companies.

So the questions that I've been asked to address, not necessarily answer but bring up points about, are what magnitude of difference between subgroups is clinically important to warrant a stratified or adjusted reference database? For which covariates are covariate-specific reference limits needed? I'm trying to change my terminology here. How should individuals be selected, with respect to the covariates, to construct the reference database? So these are some of the issues that I will be addressing in my talk.

So there are several candidate covariates for stratification and adjustment, which many of the previous speakers have mentioned. And they're candidates because there's consistent evidence -- from histology, from previous clinical studies -- that age, race, optic disc size, image quality, axial length, and refractive error do influence the structural measurements

from imaging instruments.

So we've seen variations of this slide describing the normative databases. Reference databases. Sorry. But I want to focus on how they are currently adjusting their databases for some of these covariates.

So all of the databases are corrected or adjusted for age. Some of them also adjust for disc area, at least for some of their parameters, particularly the optic nerve head and the nerve fiber layer. Some adjust also for signal strength. And in the case of age, if you're looking at the sectoral values in the normative databases, some instruments adjust for age in all of the sectors and adjust for other covariates in all sectors, even though that covariate might not be statistically significant in all of the sectors. But one does not adjust for age in all of the sectors.

So what is the evidence, for example, for age and why do we need to adjust for it?

I'm going to be showing a lot of examples from the RTVue and Cirrus, just because there's more information about those databases available.

So in this case you can see there's a lot of scatter, and the R-squared is a measure of how much age explains the scatter in nerve fiber layer thickness in these databases. So the R-squares here are 7% or 11%. That's not a whole lot of variability that age explains. And the rate of decrease, actually, in these two databases was relatively similar.

But if you look at the age range that we're looking at, I'm going to say from about 40 to about 80, there's about a 10  $\mu$  difference between the mean values at 40 years old and at 80 years old. So one could argue that, well, 10  $\mu$  over four decades, do we really need to adjust for age? That's a question.

The other issue is that the rate varies by sector and baseline thickness. This is a study in Hong Kong, so it's a Chinese population, more myopic than perhaps the other databases. But it shows that the average thickness, the rate of change per year was .52, where in some of the sectors it's much larger, and that this also varies by what the nerve fiber layer thickness started with at baseline. So this is a longitudinal study as opposed to a cross-sectional study.

So I'm going to be presenting arguments for and against adjusting for age.

So, for: There are small significant differences that we know from histology, from other imaging, and also from the data that we're using for the reference databases. Age does explain some important variability. You can argue whether the R-squared is relatively small or large, or whatever, in rim area, for example. It is smaller than for the nerve fiber layer. The rate of nerve fiber layer decrease is much larger in the sectors.

So if we're thinking that we see glaucomatous damage in certain sectors earlier, then perhaps -- especially in the sectors we should be

adjusting, and it may help differentiate between age-related and glaucoma change. I think, as these instruments become more sensitive, we're going to be seeing and finding statistically significant changes because the variability, as we heard earlier, is relatively small. And whether these changes are related to age or to the disease, I think the clinicians need some help here. So perhaps if we adjust for age, there might be some benefit there.

Against: Well, there are small differences. Between 40 and 80 years old, there's only 10  $\mu$ . Is that enough to adjust for? Well, maybe that's a good question for the panel later. And also there's large variation by sector and by baseline thickness.

What about race? Well, we know there are a lot of differences in optic nerve topography by race. And I'll show you this example for disc area with the African, European, Hispanic, Indian, and Japanese; and the African and Hispanic have larger disc sizes than the other races. And this particular is from the RTVue normative or reference database.

So what the normative databases often do is, they will adjust for disc area, but we still see small significant differences by race after adjusting for disc area and age. However, what is the magnitude of this difference? Is it important for a stratified normative database to be available? Well, the largest difference in rim area between the races with the RTVue was .21 mm<sup>2</sup> and that was between the Hispanic and the Indian databases. And with the Cirrus, it was even smaller, and that was between

the Hispanic, who had larger rim areas compared to the European. Is this difference large enough?

The other issue is how does the type of normative database affect the diagnostic performance of these instruments? So this was an interesting study that used the normative database from the RTVue, but only used the African descent (AD) or the European descent population in those databases, and they recruited African American glaucoma patients and European-descent glaucoma patients.

So this is the area under the ROC curve, and the point here is, one line is for the European and one line is for the African, if that label isn't clear. But the point is that the diagnostic performance is similar, regardless of whether the glaucoma patient was African descent or European descent.

The study went even further, and what they did was, they compared the African descent Europeans, which is the three bars on the left, to three separate normative databases -- the one I just showed you that combined the African and Europeans -- and they compared it, in blue, to the African-only and, in red, to the European. Similarly with the African descent. So you can see that the diagnostic performance was virtually identical, regardless of what normative database was used in comparison.

So should we stratify by race? Well, here are some for and against arguments. Well, there's clear evidence that race does matter. There are differences. And misclassification of individuals will be reduced. But

there are several or numerous arguments against. How do we define the groups? There are a lot of mixed races. There's large variation within groups. What is a Hispanic? Is it a Mexican, a South American? African, Asian. Will we need separate Asian databases for all the different types of populations?

There's evidence that I showed you that the overall diagnostic accuracy is similar, regardless of which norms you're comparing to. And there are other parameters, such as disc area, that explain a lot of the variation that we see by race. And, of course, there's added expense of sample size sites for constructing a race-stratified database.

So what is the evidence for optic disc size? Well, there are strong associations of optic nerve head parameters with disc area. Nerve fiber layer is the small one with an R-squared of about 5%. Cup area, 42% of the variability is explained by disc area. So that is a strong association.

The strength of the disc area association can vary by instrument. These are both vertical cup/disc ratio in the RTVue. In this particular case, the R-squared was about 20%, but that's still relatively large. And the Cirrus was over 50%.

So should we adjust for optic disc size? Well, it does explain a large proportion of the variability in optic nerve head parameters and explains much of the variation by race without the need for race-specific databases. And it's readily available because it's measured in most spectral domain OCT instruments.

And the arguments against: Well, the software delineation of this margin, you might need to take a disc scan as opposed to just a macular scan or something, if you need to adjust for it. And the definition of disc margin is changing, but the instruments are getting increasingly good at identifying a stable disc margin automatically with these images.

What about signal strength and axial length? They're statistically significant, but the R-squared is much smaller, mostly less than 7% or 4%. So one could say, once again the evidence for, there's consistent, yet weak, association. But I guess it really does explain a relatively small proportion of the variation in the nerve fiber layer and optic nerve head parameters. And I am focusing on those two because there's a lot more data. But all of the thoughts about this will apply to the ganglion cell normative databases.

To summarize, the importance of a covariate depends on what is measured. Age is more important in the nerve fiber layer than optic nerve head. Disc area is more important in optic nerve head databases. The magnitude or difference or strength of the association, once again, can be statistically significant, but weakly associated. The importance also depends on whether another covariate explains some of the heterogeneity. Disc area and race, signal strength and age, perhaps, can vary by segmentation or instrument, and it can vary by sector. Do we need to be consistent and adjust, regardless of whether each sector has a statistical association or not?

So back to the questions that I was asked to address. What magnitude of difference between subgroups is clinically important? Well, it's difficult to set a specific magnitude, but perhaps a minimum R-squared or some clinically important difference that we would need to come to consensus on should be required. Once again, what's clinically important can be different from what's statistically significant.

Which covariates are covariate-specific limits needed? Well, the strongest evidence is for age in the nerve fiber layer and disc size for optic nerve head parameters.

And how should individuals be selected with respect to the covariates? Well, I think a large range relevant to the disease. You want older ages well represented for these databases and perhaps a variety of races.

These, hopefully, we'll discuss further in the panel.

Thank you.

(Applause.)

DR. SHAN: Thank you, Linda.

Our next speaker is Richard Lee from Bascom Palmer, and Richard will give us a "Review of How Normative Database Information is Displayed by Software/Printouts and Described in Labeling for Various Models."

DR. LEE: Well, I want to thank the organizers for inviting me to

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

this. I think I've got the easiest part. I'm just going to show a bunch of pretty pictures. No numbers. Probably the most I was capable of doing.

I'm going to talk about the normative data and how it's presented and how really, from a clinical viewpoint, we really use this. It's one thing to have a car engine, but you've got to talk about the driver who's going to be putting that car through the loops.

And I don't have any particular financial disclosures myself, although the institute does receive some support from industry.

So what is clinically useful information? I mean, we talked about how there were all of these different parameters, and some of these have validation, as we've talked about, and some of these don't, really. It's not really clear what are each and every one of these bits, and there are many, many more. Once you have huge datasets, you can compile them in any which way you want. The question is, is that data actually relevant?

So a simple example is, you have a fork for everything, you have a knife for everything, you have a cup for everything, and they're all useful. But what really is it that you really need to really go through a meal, okay? And for me, I'm a simple guy. A spork is great. It does the fork and the spoon. This Splayd, I don't know if you know about it. It's 18/8 stainless steel. It's Australian, of course. It is a knife on the side. Okay, it's a knife on the side, it's a fork in the middle, and it's a spoon, all in one. You can just sit there. And they're sold at nice stores. They're not cheap. Okay, this works

fine and it tells me everything I need to do to get my meal through.

So we really need to talk about something in between all of this, where we've got all of the tools and instruments, but we have to identify what we really need that's going to be helpful for the patients. It's another thing to sit there and go through the statistics and how reproducible and whatever. But what is it that we really want and what we really need?

So what information is going to help us really go from normal, or the reference space, to glaucoma, disease? And in between that somewhere is what is abnormal, because something that is abnormal is not necessarily diseased.

So just to give you an example of what I'm talking about here, here's an optic nerve head, and what we want to do is we want to really define this population so we can ask when this population is really not this population, but this population, which is not just abnormal but also diseased.

So how can we similarly acquire data to be presented usefully? Because many of these machines are really all deriving similar data. All right. It's just a question of what kind of segmentation algorithm are they using, what flavor of that piece of food you're having is present.

And I'm going to go through some of the different formats, but you'll see that they're recurring themes, because there are certain things the clinician looks at and says this is what I need to help me make a decision.

So if you look at them, you'll usually get a nerve fiber layer heat

map. You might get a TSNIT map. Here's a deviation map. And then here, of course, are some of the maps that show you the segmentation so you can tell where there might be areas of segmentation failure. Sometimes you can see that when you look at the heat map. And, of course, there are some baseline overall statistics.

And here, if you look at a different format in RTVue, yes, instead of 12 clock-hours, you have 16 slices of pie. You have a TSNIT map. You also have summary data here. If you look at the Spectralis, very similar, right? I mean, you see that they're all -- everybody has to have something that looks their own, right? I mean, if you go out there and you look at a car, all cars have four wheels and whatever, but it's the skin that they have that tells you the difference between one car from another, right? They're all going to be a V6 or a V4 or whatever it is.

But when you look at these printouts, one has to be cognizant of what the data is really telling you, and is that dataset from one platform to another the same. And that's what we're really getting at, right, the reproducibility, the statistical basis for some of these things, and propriety algorithms that you don't really quite understand, but they'll present it to you in some way that hopefully is going to be clinically meaningful.

And here's one that many of my Latin patients will bring in that's not here in the United States. But, you know, they add something, for example, here. They have the disc damage likelihood scale from

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

George Spaeth. I mean, I don't really know how the computer really draws this circle to derive a score. But these are the things that I think the FDA plays a role in, in helping guide some of this, so that some people have some sense of "can this be data that we can trust?" So we can't have the Wild West, but at the same time we don't stifle innovation.

So you have the same kind of a printout. You know, this is a Polish company with ties to Canon. And Latin America, like I said, I get a lot of these scans coming from Latin America because the platform is relatively inexpensive. But yeah, you're seeing heat maps, your deviation maps, your TSNIT. This was a little extra, you know, gimmick to bring you in to something different. There's a value added. And then here are the same parameters that you see with cup/disc ratio and whatnot.

But even within the same platform, you can look at the data in many different ways. So here are two spectral domain OCTs from the same company, and you can see there are heat maps; there are your TSNIT maps. But here, instead of having your segmentation in this orientation here, you can have it done this way. There are many ways to slice the pie that still ends up giving you what you need.

And if you look back at a more legacy platform for the same company, you'll see here this table here, for example. I mean, the only thing I really used out of that was the average nerve fiber layer thickness. I mean, I don't know if anybody used the rest of this other data for any clinical decision

making.

Having said that, I know that whatever company puts out their data sheets put a lot of time into it, because on a single sheet of paper you've got to have everything. So real estate is really important, and I think it's very critical that the companies come together to present the most important clinically useful data for us to make decisions, as clinicians, in the end.

So if you look, for example, here, we talked about nerve fiber layer. Well, another way to look at this is to look at the optic nerve head in conjunction with the nerve fiber layer. And then you get into all kinds of other different statistical issues.

One of the really important things for the clinician is when you're in your clinic and you get a printout from your technician. When I sit there and actually go through this, I couldn't understand why sometimes we'd get this particular printout when I used this platform. Well, it turns out, if you're actually there with a technician and doing a live scan, you can actually set this so, if there's segmentation failure, you can actually change the segment and manually segment this and then print that out so you know what the segmentation is, so you know what the data is that you're getting back. I didn't know that until I went down there and said I don't want this printout. Somebody actually sent up the wrong printout.

So there are different printouts that give you information, but there are also things in the back end that we don't understand. And we

talked about operator issues with regard to variability and reproducibility.

There are a lot of other things that are going on, because we have technologies that are like a Ferrari and we all are driving it like a Volkswagen because we don't exactly know what's all under the hood because, at the end-user side, we just need it to get us somewhere.

So here again, if you look at another platform, here you can get your optic nerve head analysis, you can get your datasets. And when you look at different platforms, you get like Moorfields Regression Analysis, you get all of these different scales. But at the end of the day the question is, is this analysis going to tip me over when you put it all together into making a clinical decision?

Now, there are three slides here that I was asked to kind of put in per some of the questions that the FDA had, and that had to do with this reproducibility issue. And hopefully we'll talk about this more when we have the panel discussion.

So, for example, just to give you an idea, within a platform, if there are things that are particular off of the normative database in terms of reproducibility limits, there's actually a little marker that tells you something is going on. And so here, if you look at this, it's all green, but there are some questionable areas down here. So there's a difference when you're doing a larger quadrant analysis versus a smaller segmental pie.

Well, if you actually look into this further, you can actually get

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

these printouts, which most of us don't. I actually wasn't even aware of this until I was asked the question, what do you think of this request from the FDA to look at the reproducibility limits?

And so what this does is it actually adds and subtracts the differences, the plus/minus error, basically. Okay. So it's a linear add and subtract. And the question is, does that tip you over into a different color, into a different statistical band? So something may be yellow, but on one hand it would be green; on the other hand it would be red, all right? But that data has already been there.

So some of these things that are being requested aren't really clear that they're going to make a huge difference for the clinician, because I'm not going to waste a tree to get this printout when I'm in the clinic, because the fact of the matter, it's a plus/minus, and I can do it in my head, for one thing. But the other thing is the plus/minus. Here's the key point. The plus/minus changes depending on where you are in your circle scan, right? So there are simple ways that you can actually approach that information.

So, for example, what I do is, when I look at this heat map, I look at how dense this heat map is, and this tells me I'm doing pretty well. If it's a lighter color, I know I'm going closer to the bottom of my limits. And, of course, one way you could present all of that is, when you look at your TSNIT map, make your line thicker or thinner, depending on what your variability

limits are without having to have a whole new sheet of printouts, one for the right eye and one for the left eye.

So there has to be some thought, I think, into how this data is presented, but on a single sheet in a simple way that we can quickly recognize what's going on, because, as I pointed out, this data is presented very similarly for multiple different platforms. Like here it's for the GDx. It's the same idea, right? You have your deviation maps and you're looking at your heat maps with the nerve fiber layer. And, of course, this is going to get even more complex when we talk about doing ganglion cell complex analyses.

So when you talk about what I would consider as a clinician, in terms of these printouts and how it's going to help me, especially with regard to the statistical analysis that underlies this data that allows me to make a decision, I look at the nerve. There's nothing that still beats looking at the nerve head because you can't catch things like optic nerve head hemorrhages and other things with this analysis. I look at the visual field to get a sense of the function, and then I look at the OCT and taking into account some of these caveats about reproducibility, sensitivity, and whatnot. Some companies are doing this by integrating some of their data with regard to OCT and functional data, such as visual field.

And as I pointed out, the simple thing that I usually do -- because the thing is, we have all of this data. Ultimately, it is the clinician that takes all of this into account, into what is actually a disease process and

what is not a disease process.

And so it's the integration of looking at this, doing the pattern recognition, looking at the data here, looking at the function, putting it all together and then making a clinical decision. So this is not something that's taken in isolation with just one test, just as we do for many of our other clinical entities.

So current imaging technologies allow for the analysis of multiple parameters of the optic nerve and the nerve fiber layer. The clinical usefulness of some of the data is not really known or it's not really clear. And the validated clinical data is important for discriminating normal from glaucoma and progression in disease. That could be highlighted later on for some of the discussions.

Thank you.

(Applause.)

DR. SHAN: Thank you, Richard.

Our next speaker is Gustavo de Moraes, who is from NYU and New York Eye and Ear, and he will talk about "How does Construction of and Statistical Modeling within OCT Normative Databases Compare with Standard Automated Perimetry Databases?"

DR. DE MORAES: Thank you, everyone, for the opportunity.

These are my disclosures.

And I'd like to start my talk with a message, a wakeup message,

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

that even though we're talking about OCT and structural testing, the first consensus of the World Glaucoma Association said that all patients with or suspected of having glaucoma should be diagnosed and followed using both structural and functional tests. We know that it's not always that we find an agreement between structural and functional tests, but when it happens, it increases our likelihood that there is disease or there is progression.

The fact that the development of OCT devices has been so fast and so dramatic in the past years, which has not been followed with the same velocity with perimetry, gives us the importance of this talk, which is to look at what points of these new technologies, these new OCT devices, need to be paid attention when comparing with the visual fields. A lot of the times what we call disconnects or disagreements may not be actually disagreements.

So let's start talking about perimetry. So it's the general concept that the AGFA (ph.) contains different normative databases that provide data for a statistical comparison of how patients' visual field results compare to an age-matched control. And this is done both cross-sectionally for diagnosis and longitudinally for progression. So let's start with the cross-sectional analysis.

I'll not go into details about how the database was developed. You can find it in different textbooks. But the same group of patients was tested to develop the SITA and the SITA SWAP normative database. And the SITA normative database I'd like to highlight here. Collected data went into

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

the creation of two databases, SITA Standard and the SITA Fast database. So just keep that in mind.

The data was collected from a prospective, randomized, multicenter study with healthy subjects between 17 and 89, which is similar to the OCT inclusion criteria you just heard. And, again, no history of eye disease and followed by comprehensive ophthalmic examination.

I'll not go into details about the inclusion criteria, but they're very similar to the ones presented for OCT.

I would like to highlight that IOP was used as a criterion, as well as best correct visual acuity, refracted error, suspicion of pathologic optic discs, so there was some type of structure evaluation, and visual field defects or suspicious visual field defects were also considered and used as exclusion criteria.

And the demographics of this SITA normative database, they entered 422 eligible subjects with a mean age of 53. And what's interesting and relevant to this talk here is that information on subject ethnicity was not collected as part of the database selection, even though the data was collected from centers in Asia, North America, and Europe. So we're not sure if we're comparing a specific subject with a specific ethnic group.

This is a study from the 1980s that investigated the normal variability in static perimetry, showing some concepts that we're familiar with, such as that variability changes according to eccentricity and, moreover,

variability changes according to severity of the disease.

And why is it relevant in terms of OCT and the new technology we have today? We're not just only looking at the optic nerve and the area around the optic nerve, but we're giving more attention now to the macula, as we can get not only measurements of the retinal nerve fiber layer, but retinal ganglion cell.

And this information is very important because the retinal ganglion cell layer in the area of normal or typical OCT scan of the macula will involve an area that is a very small area of your usual 24-2. So the entire area that your OCT is imaging actually occurs approximately 16 points of your 24-2. And whereas if you use a 10-2, we now have approximately 69 points representing the same area. So the point here is that we may not be detecting damage in the macula using 24-2, and this could be one of those causes of disconnect which OCT may help us circumvent in the near future.

This is a real patient of ours that had a normal 24-2, but when you looked at a 10-3, saw these clear arcuate defects. And if you look here in the red or pink circles here, the points from the 24-2 would fall in areas outside the main area of the defect. So the defect would have been missed here.

Also, one of the intents of these new technologies is to combine the evaluation of the optic nerve and the macula into a single scan. And because of that, it's imperative to combine our information from visual

fields to that information, not only what the visual field and the structural or functional maps correlating optic disc and visual field exists, but also some correlations between the macula and the visual field and the macula and the optic disc.

So this is just showing an example from our group, courtesy of Dr. Don Hood. This is a patient. On the left here, what you see are the retinal ganglion cell measurements from patients going from controls to mild to moderate glaucoma, and on the right is a subtraction from normals. And this is a patient that clearly appears to have some loss in this part of the field. But, again, because there's not a significant representation of the points on a 24-2 here, you may miss a lot of these defects.

This is another example. Now, this is the part that I think an OCT normative database, in combination with visual field, will help us improve the agreement between structure and function. We can now, with significant accuracy, overlay or superimpose visual field information to retinal ganglion cell and retinal nerve fiber layer information, which again increases the likelihood of finding a defect. Sometimes a defect may not be statistically significant using a standard 1% or 5%. But if you looked in a spectrum where you're using a continuous probability, you'd probably find that that area may actually correspond to true disease.

So this is showing an example of the points that were abnormal and the 10-2 actually corresponded to this area here. You can superimpose.

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

And this is our impression, that this may help improve the diagnostic ability in glaucoma.

These are some examples of several cases that it's not like a clear defect that you would just base on the field. But once you look at the RNFL images, you can see that there is some scatter that matches well. So one case that you usually consider a case that needs to be repeated, with a combination of an OCT scan, you may increase your certainty of the diagnosis.

This is another SITA example showing an area that apparently would not reach statistical significance at our usual 1% and 5% levels. But if compared with an area that also has some damage -- this is so it matches the visual field and the optic disc -- that will make us more certain about the diagnosis.

And the importance of this is, as I stated, are represented at ARVO this year, is that there's probably -- the defects in the central field are probably as often as defects seen on 24-2s. What it means is that if you don't perform 10-2s in a lot of patients that have apparently normal central 24-2s, you may be missing some defects. And for the purpose of creating a database for OCT, you may be including a lot of glaucomatous patients in your normative database.

And this is actually a study from long ago. A few people have probably read it, but it shows that up to nearly as many of 30% of the cases

have abnormalities on a 10-2.

Now going on to perimetry again, looking at longitudinal analysis, there was very little in the discussions here on progression. But the GPA has a progression analysis that uses patients within a different spectrum of glaucoma followed for short intervals, so that the inter-test variability would be defined and progression would be defined based on that, which they have this criteria which I'll not go into detail.

Again, this is not representative of ethnic groups. It's more related with age of the population. And in both cases, both for cross-sectional and longitudinal analysis, the standard cutoff values of 10%, 5%, 2%, and 1% are used.

And this gets to the point I want to specify. I'm not going to go over the optical coherence tomography criteria. The previous speakers gave an excellent talk on that. But I want to highlight some important points.

There's nothing available today with the commercially available devices in terms of progression. At least this analysis. They don't have this analysis of inter-test variability based on severity of disease. We know that when you use linear regression to look at progression, it violates many of the principles that are supposed to be respected, such as that the variations are seen throughout the follow-up. You know that as this disease gets more severe, you see more failures in the algorithm. So there's an imperative need of learning what perimetry did to develop these longitudinal follow-up

databases.

And the implications. I just want to finalize showing the map and the model that Dr. Hood developed of structure and function. And the point that I want to make clear here on this map is that if you're going to detect glaucomatous first based on visual field of OCT, it depends a lot on where in this variability, this range of variability, where you are. So if you're someone that's closer to the lower bound of normality in the visual field part, you're mostly likely going to -- that goes with fields before OCT. This is just one of the hypotheses.

And just some examples from the literature. This was shown recently, the ADAGES study, in which individuals of African ancestry had worse visual field sensitivities than whites, and they argue that it can be a problem of the database. Why not the same thing is happening with OCT? And the same thing happened when looking at structural tests. So why not an inappropriate database for structure may be the cause of these differences?

And this is the last slide. In the Ocular Hypertension Treatment Study, if you need to be reminded, black ethnicity was a risk factor in the univariate model, and who knows if the cause of that was also that these probably were at the lower boundaries of normality in using our current databases?

Again, this is what I want to show. These are the differences of

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

all the points we need to address when comparing the two modalities. And, again, we should follow the patients with a combination of tests, especially if you can use continuous probabilities instead of the 1%, 2%, or 3%, as was recently shown in a nice study from Wall and colleagues.

And thank you.

(Applause.)

DR. SHAN: Thank you, Gustavo.

Our last speaker for this session is Kristen Meier, who's from the FDA, and she will tell us about "Normative Database Construction, Reliability, and Usability: Considerations in Premarket Review."

DR. KAHOOK: It's not working. Is Richard Lee in the room? Do you have a Splayd with you? No? A Splayd, anybody?

DR. MEIER: Thank you. Let me just acknowledge again Dr. Lee Kramm, who provided significant contributions to just this whole workshop and to my talk, and also to my understanding of OCT and this technology; also to Rona Tang, who helped a lot with the statistical content of this talk.

So, first of all, again, a lot of you have worked with this technology. I was new to this technology. There are lots of different parameters that are put out by this device, so there's an awful lot of data. Let me just focus for now on one parameter and just, for example, RNFL thickness.

There are two types of databases, or at least two types of databases you might think about. One is the cross-sectional reference or normative database, which is what we're talking about here. And in that database you're taking results from many individuals, each contributing a single result. In contrast, you would have a person-specific database, and that's where we would look at results over time on the same individual. And those types of databases, I think, are very useful for patient monitoring.

I want to just add a slide because there was some discussion about this, this morning. What we're not talking about here today, I think, is when you're taking multiple parameters and through some software algorithm optimizing or training that algorithm to predict somebody's outcome. Those kinds of algorithms are actually a lot more complex; and we encourage, at FDA, if that's what you're interested in working on, that you come in and talk with us. There are currently no cleared OCT devices that actually have those kinds of algorithms in them.

Again, a lot of my points actually have been touched on through the earlier speakers. I also agree that normative is a potentially very misleading term. It can have several different connotations from the statistical sense of the distribution of results that follow a Gaussian or a normal distribution, but that's not necessarily the same as being typical of a group of individuals. It could mean the absence of only the condition of interest. It could also mean the absence of all possible diseases. So it's not

really a clearly defined term.

I believe, as well, that reference is a better term, that it prevents some ambiguities. And I think it also forces the issue that you really need to know what constituted that reference group. When you say normal, a lot of people can have misunderstandings about what that means. Now, I agree that that's probably a better term, but in my talk, I will use normal because that's what I think the community is most used to.

I'm going to skip a few slides just because I think some of the points were already covered here.

Again, typically, the way that normative databases are used is by comparing results to that database. And as already discussed, there are two general types of approaches of the use. I'll call it the normal limit approach, where you have a dichotomy and you flag or color code or somehow identify somebody as exceeding that limit or not. A lot of times people are using the fifth and one percentiles, but honestly, there's no magical reason why those need to be the points that people are using.

Alternatively, there's the percentile approach, where -- and I shouldn't say alternatively. Actually, I think, in addition. And perhaps you could also report actual percentile to be clear exactly where that result lies with respect to the database.

For a single parameter, by design, then, 5% of the time a result from a normal person is going to be outside that normal limit. But as you saw

from the first slide, there isn't just one single parameter that we're looking at. We're looking at lots, dozens, of parameters.

So I don't know if many people think about it -- it was brought up earlier -- that if you're comparing multiple parameters each to, say, a 5% limit, the chance that any one of those is going to show up outside the limits is quite high. If you had 10 parameters, that percent is more like 40% if they were independent, which they're probably not all independent, but it's still quite high. If you had 20 parameters, that percentage could be as high as 64% of a normal eye, having one of those parameters outside. So that can explain easily why you might be getting some red spots on your images, just by chance alone.

Again, just to reiterate, the percentile approach is an approach that's familiar. It's used in pediatric height/weight growth charts. And I think it's nice because it avoids the dichotomy of normal versus outside normal.

I think that by using normal limits, it may get misinterpreted as a clinical decision limit. And by that, I mean a limit that discriminates between states of health. To really do that, you need additional information. You need distribution results in disease or undifferentiated subjects. I think the normal limits is a great starting place to help you figure out where those clinical decision limits are. But I think we can't stop there, at just looking at normals, and seeing where these limits are.

Whether you're using a percentile method or the normal limit

approach, they are both being compared to a well-characterized normative database or reference database, and that is why one of our first questions focuses on what type of people should be in that database, what's the right reference group.

Again, already discussed. When results differ across different covariates -- I'm using the statistical lingo here -- different age groups or ethnic groups or different levels of image quality, those normal limits may need to be covariate specific or stratified or adjusted for those potential differences.

A hypothetical extreme example here. Suppose this were a distribution of results, of some thickness results, say, from a normal population, you might set the 5% limit down at this line here. Well, if you didn't look at the individual subgroups, you might not appreciate that there are huge differences in subgroups.

If you look at the next slide here, a person, say, where this blue arrow falls, would be way outside for Subgroup 1, to the other end of Subgroup 3, but sort of in the pack with Group 2. So it really matters. We need to figure out what kinds of covariates and what subgroups might be important.

I also believe that this was already mentioned, but I want to reiterate, as a statistician, that statistical significance alone should not be the criteria that's used to decide. A lot of people feel that it's more objective, but

in fact, it's not. You can manipulate statistical significance by sample sizes and things. So that should not be the main criteria for deciding what factors should be considered for adjusting for covariates. I think it's important to consider the magnitude, the direction, and the clinical significance of potential differences.

We also look, at FDA, whether the covariate has been identified or established in the scientific literature and whether that covariate is used in similar devices already on the market. Questions 2 and 3, which you have copies of, will get into this hopefully in the discussion period.

For those of you that do database construction, I just have some points outlined in the next two slides. And let me just reiterate for those that don't know, all the slides from this workshop are actually going to be made available to folks. I know a lot of these slides we're going through very quickly. So all of these slides will be available, I'm not sure, on the AGS or FDA website. But, again, I think most folks are familiar with these concepts, or if not, again, you can refer to these slides afterwards.

So here are some points to consider. I think it's important to prospectively plan. When you are constructing a database, we're interested in good subject selection criteria, clinical workup, recruitment procedures, testing protocols, recording image quality. If there are any statisticians in the room, we're interested in some additional things here. I think the most important thing on this slide is that you want to consider the need for

covariate-specific limits and be sure that you measure all of that information.

As we know, device models are evolving, and there is interest in transferring databases from one model to another. These are expensive studies to do, I'm sure, and folks would like to be able to transfer databases.

To be able to transfer a database as is, I think that you really need interchangeability results, which is not necessarily the same as substantial equivalence. That's a different type of a paradigm. I do think, however, that it's possible to transfer databases with an appropriate calibration that might be estimated between models.

And although this is not a lab device, the lab community has thought long and hard about a lot of these issues actually, and FDA uses a lot of Clinical and Laboratory Standards Institute standards, and in my references in here are just two standards. But, again, although they're written in lab lingo, really, the statistical concepts are directly applicable to OCT devices.

Database reliability is something FDA is very interested in. And, again, we've talked about variation throughout this workshop already. There's variation in the measurement, and there's also variation in the estimation of the normal limit. And that can vary for many reasons, a few of which are listed here.

It depends on the actual characteristics of the subjects in the database or the spectrum of subjects in the database. It depends on the sampling variability. Is your database 100 people, 300 people?

It varies depending on the statistical model you use. We've heard some nice talks on the statistical models that are used. It's important to verify the adequacy or the fit of any models that you're using.

And it also is impacted by the actual variability of the measurement itself, which was this morning's discussion, where we talked about repeatability and reproducibility.

What should be conveyed to the user? We believe it's important to convey several things. We believe it's important to have a very good characterization of that normative database population. I believe it's also important to characterize the variability in both the estimation of the normal limits and of the measurement. The question, of course, is how to do that.

Just a few points here for things that I think are important for characterizing a normative database, starting with what is the definition and clinical workup that was used, what were the recruitment inclusion/exclusion criteria, testing protocols, how scans were deemed acceptable for use, and whether there are any limitations of use that would be identified.

In the user's manual, which I don't know how many folks actually read those, or labeling -- but we do. I know manufacturers spend a lot of time on those user manuals. FDA spends a lot of time reviewing user manuals and being sure that appropriate information, performance information, is in there and also any disclaimers that a user might need to be

aware about use of the information.

As I mentioned, normal limits, themselves, are subject to variation. I have an example displayed here for how one might convey this information to FDA. Whether this is useful to the user or how this might be conveyed to a user is one of our questions.

Here, as you know, when you have covariate adjustments, you now have to specify some covariates to be able to give examples of what some of those limits might be.

In terms of variability, I think it's important to see -- if we can't even discriminate our estimates of the first percentile and the fifth percentile, then it's not clear that having color codes that distinguish these is useful. So I think the variability plays into this, and if we can think about how best to do that and get some ideas from this workshop, that would be great.

This was brought up in Dr. Lee's presentation. Another type of variability that is conveyed is the reproducibility, which again incorporates the repeatability within a device, but also between operator and device variability.

So sometimes just an original result is reported. I think adding the percentiles and some additional information here, if you then color code in addition, you can see how the reproducibility might impact that. Now, I'm not for killing lots of trees either here, but I do think it's somehow important. I don't know the best way to do it, but it's important to convey how that

reproducibility would impact. When you say a result is 94, well, is it 94 or is it really 70? Is it 118? And in terms of percentiles, is it less than 1%? Could it be as high as 15%? Somehow conveying that kind of information, I think, would help improve the usability of these databases.

And, again, those are some of the questions that we're interested in, in Question 4.

In summary, I believe that there are several factors that we need to consider. Lots of the talks that we just heard brought up a lot of these issues. The characteristics of the database population are important. The need for covariate-specific normative limits needs to be thought through. The appropriate choice of the statistical model for estimating those limits needs to be verified. And, finally, but not least, characterizing the variability of the result and the impact of that variability on the comparison of the result to the normative database is important.

So thank you, and we're ready for questions.

(Applause.)

DR. KAHOOK: So I'd like to thank all of the speakers from what was clearly the best session of the day, I think you guys would all agree, and have all the speakers come up to the table here so we can start some question and answer.

I want to remind the audience that we do have open mikes, and please do come up to the microphone. And also those who are attending

on the webcast, please do send your questions, and I think it's important to all of us for you to participate.

One thing we're going to do slightly different from the earlier session is -- and this was at the request of the planning committee on both sides. I'm specifically going to go over each question. I'll then ask the panel if they have anything to add. And several of the speakers mentioned that they would like to add certain content during the panel. And then after that, I'll ask the audience if they have any questions that they would like to add. And if there is no need for further comment, then I'll move on and we'll see how many of these we can get through.

So the first question is up on the board here: What clinical work-up should be done to establish a "normal" population who can be defined as not having disease? What patient characteristics are important to represent in the definition of a normal population?

And several of the speakers touched on this topic, and it sounds like Rob Fechtner might have a little bit more to add.

DR. FECHTNER: Well, I presented an alternative viewpoint, which is that perhaps there is not a detailed work-up that needs to be done to establish a normal population, but rather, we need to get a broad population-based sample as our reference population.

And I wonder if any of the other panelists have thoughts on whether that is a useful approach for a reference database?

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

DR. KAHOOK: So none of the clinicians have anything to add. I'm going to Kristen, maybe, if she can add a little bit on this concept of -- let's call it a reference database -- and being more inclusive, looking for a more heterogeneous population within that database.

DR. MEIER: Yeah, I've kept the normal terminology here because, again, that's what is used. But I think, absolutely, we're interested in whether, in fact, normal is the right thing; even again reference, thinking broadly, what is the right comparator group. Again, you compare a result to that group. What's the right reference group?

So, again, I would be interested in some of the other speakers as well, other panelists, about what -- and you don't need to restrict to that the super-normal is the right notion. We're thinking broadly here. What is the right reference group is perhaps maybe a better way to state the question.

DR. KAHOOK: Professor Zangwill, do you have anything to add, given the content of your talk, about the different covariates that could be looked at?

DR. ZANGWILL: Yeah. Well, I think just in terms of the super-normal group, I know that often there's exclusion criteria for systemic hypertension or diabetes, and I think those kinds of things definitely -- I would argue against those kinds of restrictions, especially when we're looking at the age group that we're comparing to. So many people have those

conditions that it really is making a super. And whether you need, as well, I think, a broad range with less exclusion criteria is good.

I agree with Rob that a general population might be good. But then you get into sample size issues. And I think for that particular paradigm, to work without any kind of exclusion, you really do need a large sample. So that's kind of a cost benefit issue, that perhaps there are some exclusions, that you would want to have it less likely to have some kind of disease. But I'm open to discuss that.

DR. KAHOOK: Is there anybody from industry in the audience that would like to comment on this topic specifically, creating a normative database?

We have a taker. We'll give you a couple of seconds to work your way to the microphone, then.

MR. SAND: Thanks. I am --

DR. KAHOOK: Can you just state your name, affiliation, and any disclosures?

MR. SAND: Mike Sand (ph.). I work for Topcon.

And I could just say, I think Rob Fechtner has a great, great idea. I think it would solve a lot of problems. And especially since the disease prevalence is so low, if you do the data collection the right way, I think it would really work. So I would support it.

DR. KAHOOK: One thing that might be nice during the panel

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

discussion is to attempt to come up with some concrete action items.

And I think, Kristen, I'll ask you again. With what Linda and Rob have added, and then what Mike just commented on, any interaction that happens when a discussion is going on for creating a database, how can that be influenced by some of the content that we just discussed, as far as how to create the normative database?

DR. MEIER: Companies hopefully are listening and are here today to hear this. And we often see just what they come up with -- an initial proposal -- and then through our pre-submission process, discuss with us, before beginning to collect what data they want, that they have a plan, first; that they put forth a plan based on discussions that are here today and present that and have discussions with FDA.

DR. KAHOOK: Okay. I think Joel Schuman might have a question here.

DR. SCHUMAN: I think Doug was raised before I was.

DR. KAHOOK: I'm sorry, I didn't see you. Doug Anderson.

DR. ANDERSON: Yeah, Doug Anderson.

I'm not with industry and you asked for industry, except that I'm a consultant for Carl Zeiss. I'll disclose that.

I think something in between what Rob was saying, a super-normal database has some advantage, but then Linda brought out certain common things, and it may be the prevalence of the disease that matters. So

let me use retinal nerve fiber layer thickness as an example.

This is something which is affected by pituitary tumors, it's affected by multiple sclerosis, it's affected by glaucoma, it's affected by central retinal artery occlusion. So it's affected by a number of diseases. And I would think that in your reference database, you would not want to have individuals with that disease. You would like to know, in people who don't have those diseases, where do you stand percentile-wise?

The closer you get to a low percentile, the more you begin to think this person may have a disease, in all probabilities, and the more you think of that -- and you also have to then ask the question, which disease might this person have? And that might be based on some of the clinical findings that the person has and that they come in with.

On the other hand, if you have a disease such as, let's say, hypertension, it doesn't affect retinal nerve fiber layer thickness, then why exclude them from the normative database? They should be included because they're in the same age range as the people who have glaucoma and some of the diseases that you're interested in.

So I think it may be that for different parameters that are being measured, nerve fiber layer, retinal ganglion cell numbers and so on, that you would need to have a different reference set, I don't know, depending on what disease is under consideration.

But I don't particularly like the idea of including a lot of people

who have a disease that might affect the measurement under question included. And I know that Rob said that it had to be a low prevalence disease that was included and then automatically they would all be down in the one percentile.

DR. KAHOOK: Thanks, Dr. Anderson.

Let's see what Joel Schuman thinks about this.

DR. SCHUMAN: So I agree with Doug that you probably don't want to include the people with the disease that you're looking for in the reference that you're checking against.

And we stopped calling these people normal a long time ago, in terms of the population. We call them healthy. And so I would say, having healthy controls to compare to is probably more optimal if what you want to do is to pick up disease. It's going to give you more false positives, right, but you need to be able to separate out what the machine is giving you from the total picture of the patient.

So you make a tradeoff. And I would argue that the tradeoff should be where you would perhaps call more people outside of the normal range and not include people who might have disease, might not be healthy, in your reference set.

DR. KAHOOK: Okay. Well, let's take two more questions with the people at the microphone, but just be as brief as you can be.

DR. SINCLAIR: Steve Sinclair from Biometrics and IMIT.

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

May I respectfully disagree? I'm a retinal specialist. I am an interloper here. But more and more I am seeing eyes that are seen 10 years out after retinal detachment repair -- occlusions, diabetic retinopathy, et cetera, that have what looks to be progressive nerve fiber layer loss.

And I think that what I'm encouraging companies to do is to develop feedback databases which will feed back not just the progression of that instrument's database and measurement changes over time, but also these additional factors, such as the systemic history, et cetera. We need to build these databases and manage these complexes so we can do, really, multivariate, chronologic, individual risk predications.

DR. KAHOOK: One last question for this section, and then we'll move to the next.

MR. PATELLA: Mike Patella, Carl Zeiss.

Two observations, briefly.

One percent prevalence of disease is going to have significant effects on your 1% limits. Point 1.

Point 2: I'm in favor of excluding people with normal disease. Doug and I are in alignment again. However, I'm not in favor of excluding people who have funny-looking optic nerves or funny-looking fields.

And I'd like to correct a misapprehension about the perimetry normative database. We did not exclude people from the perimetry normative database on the basis of funny fields, unless subsequent directed

examination exposed clear disease.

DR. KAHOOK: Okay, we're going to move on to the next question here, and hopefully the FDA has some information to work with there.

So the second question: Measurements within the normal population may be heterogeneous -- as we were just discussing -- and there may be differences in mean measurements for different age groups, for different race/ethnic groups, for subjects with different image quality, or for the other covariates that were discussed.

So the first question: What magnitude of difference between subgroups is clinically important to warrant covariate-specific normal limits?

And then I'll ask the second question and we can answer both of them together: For which covariates are covariate-specific normal limits needed?

And I'll ask Linda Zangwill maybe to comment on this, if she has anything to add to what she had discussed in her talk.

DR. ZANGWILL: I don't have much to add from what I said. The evidence for race-specific normative databases, I think, is limited, and I think if you adjust for disc size, that probably will cover it.

And in terms of the covariates, the strongest is with age and disc area. The only thing I would add is that sometimes there are interactions with age and signal strength, for example. So I'm not quite clear -- perhaps

there is some literature, and people can add to my comments -- that if age is a signal strength that is magnifying the effect of age or is it really explaining part of the effect of age. And those kinds of things need to be considered.

But age and disc area really are the things that I think might warrant some adjustment.

DR. KAHOOK: I think one of the issues that came up in the earlier session was we didn't have a concrete number, a concrete 5  $\mu$  or 5%. But we recognize that the answers to these are being discussed and they're fluid.

From an FDA standpoint, Kristen, what kind of answer would you be looking for, for (a) or (b) in this case, that would help in what the regulatory process is for the clinicians and for industry?

DR. MEIER: I think Linda helped answer it, and that is if companies do have databases, I think researchers have databases for exploring. As you mentioned, the race issue is something that has been a big issue for us, that we just don't have access to a lot of data or, if we do, we can't publish on it obviously. But I would encourage manufacturers and researchers to look at those issues, look at the correlations, and see. But I think Linda actually answered in a helpful way for us.

Thank you.

DR. KAHOOK: Okay, we'll take the one question here from the audience before moving on.

MR. PATELLA: Mike Patella from Zeiss again.

Linda put forward a nice quantification of what could be important in looking at age. Ten microns effect over 40 years in age is about two steps.

If we think about borrowing an idea from Henry Jampel and Harry Quigley, a number of steps measurable for normal to blind, that's about two steps in the device that can measure about 10 steps. I'll kill for two steps, and I certainly, therefore, will go after an age correction. That would be sort of my metric.

DR. KAHOOK: Okay, I'm going to read this. Did you want to comment on that?

DR. MEIER: I guess we're interested more in new covariates because, again, I'm not sure that it's all that complex to add age. And, again, if you can have a gain with it, then why not do it? It makes sense, I think, clinically. It makes sense statistically. Even though it may not be huge, it still adds and increases the precision.

So, again, we're interested especially in other covariates maybe that we have not thought about; or how to get at the race issue, which, again, I think Linda talked about very well.

DR. KAHOOK: Okay, Question Number 3 is really something we just discussed. I'll just read it for the record: For significant covariates identified in response to Question 2, how should individuals be selected with

respect to the covariates to construct the database? Specifically, what range/spectrum of these covariates should be included in a database?

And, essentially, we've had that conversation.

So the last question on the list from the program packet, Number 4: What information related to the normative database and the comparison of an individual's result to the database should be included in the labeling, device printout, and/or software interface to improve usage of these devices and to facilitate an informed decision about how to apply the output of the device?

And I'll just go through (a), (b), and (c) here:

What does the user need to know about total variability of the measurement?

What does the user need to know about variability in the estimated normal limits?

And then: Should the device report the measurement results as a specific percentile of the normal population in addition to designating into a classification?

And you know, Richard Lee, you made one comment that caught my attention, and that is the idea of real estate on a printout.

If you have just a limited amount of real estate and the sheets are pretty full as they are, how much should we ask the companies to reach back into the user's manual or to some of the material that was used for FDA

approval and include that onto the real estate that is limited?

DR. LEE: So there are a couple of issues with regard to presenting the data and then what is the meaning of that data for the practice and clinician.

One thing is, when we're talking about -- for example, I touched upon the issue that Dr. Meier brought up, that I put up some extra slides on, regarding the reproducibility limits. And the issue there is, is it plus/minus this difference? But it's not that simple, it's not a linear thing. We're talking about a tissue here.

If we're talking about if the floor is 50  $\mu$ , in terms of where there's no more nerve fiber layer tissue, and you're talking about a decrease from 70  $\mu$  to 60  $\mu$ , that's a 50% loss, all right? So that's different there. In terms of the variability, that is huge, possibly, as clinical significance for that patient, whereas if you're going from 140 to 135, not a big clinical difference.

So where this difference matters depends on the disease process, for one thing. Now we're talking outside the normal because that's where we're going to apply it, where we need the normal to have a barometer.

So I think we need to be cognizant of that because it's not a linear target. We keep on talking about this number. It's not a linear number.

The same thing like IOL calculations for the clinicians. You

know, you determine your axial length or whatever, a certain axial length is linear. Beyond that you look at other formulas.

And if we're looking at a single printout, like I said, on the TSNIT map for those same two tables I showed you, well, if you made that TSNIT map in terms of that linear graph, if it was wider or thicker, depending on -- because each clock-hour is different in terms of its reproducibility limit.

DR. KAHOOK: Yeah, Richard, there is something about that.

But I just want to give Kouros a couple of -- is Kouros up here and I just can't see him? Nice job trying to hide there.

You mentioned something about sliding scale significant levels being part of the output from the devices. Can you see that fitting onto the printout as part of the real estate, instead of the way things are presented?

DR. NOURI-MAHDAVI: Well, I think there's a point to doing that because 5%, as we look at it, or 1% -- you know, you could be, like, at 6% and be considered normal and fall into the green. It just needs another session of imaging to make you fall in the yellow area and vice versa.

So I think if you know from the beginning that you're kind of -- the patient is kind of close in that particular region or, globally speaking, close to the 5%, where you would think it's statistically important, significant, then it's helpful.

And then you could also -- if you have the information, you can actually compare the percentile around the clock on the TSNIT curve and look

at the percentiles globally and regionally and kind of come to a conclusion, use your basic pattern recognition in your brain to come to a conclusion that this is possibly abnormal or things are going down, although things look okay and are all green.

DR. KAHOOK: David, did you have a comment about this?

David Greenfield.

DR. GREENFIELD: Yeah, I wanted to point out a thing that's probably obvious to all of us, and that is the user is the physician and the stakeholder is the patient, and what's key to both of those individuals is knowledge. And if the user has a very fundamental understanding of what's in the normative database, they're more likely to make intelligent decisions about the information that they're interpreting.

And as a corollary to that, I think it's important, then, to expand all of the covariates that could potentially affect an incorrect classification, like expanding, and making sure that you have a relatively large group of elderly myopes that would give you a false reading, a false red reading, for example.

The second is, the repeatability of the disease classification is very important. If a patient is classified as abnormal, how likely is that to be a repeatable classification? You know, something that's highly repeatable is useful information that should be on a printout, I think.

DR. KAHOOK: Okay. So just to sort of close out --

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

DR. FECHTNER: One quick response to David --

DR. KAHOOK: Go ahead.

DR. FECHTNER: -- which is if red is not abnormal but it's a statistical limit, and that's the understanding of the user, it changes that dynamic a bit. So there's nothing abnormal about a red in a myope. It puts them in the 1 percentile. And I think we need to be cautious as we go forward with this, about what our language is. And I agree with you completely.

DR. GREENFIELD: Or it just means it doesn't match the reference database, and maybe expanding the reference database more appropriately might eliminate a lot of that.

DR. KAHOOK: This is a bit of a moderator's nightmare. There are four lights on with the microphones here on the stage.

DR. LEE: African Americans are a great example of that, where I've had some people come in, entire families with glaucoma, and you ask them to bring in their photos and everything else and their visual fields, and you realize they're just as illogically cupped. And, you know, every once in while the machine will catch something that looks abnormal.

And one last thing about the presentation of data. There's a three-volume set by a guy named Edward Tufte, T-u-f-t-e. You've seen it?

DR. MEIER: Um-hum.

DR. LEE: It is one of the best volumes out there on how to

present data. It's a beautiful book to read. I bought it way before my interest in any of this. It's a beautiful set of books that tell you how people have looked at data and how to present it in a most remarkable manner. It's almost like an art coffee table book. But Edward Tufte.

DR. KAHOOK: I think if Kristen doesn't have it, we can all chip in and get her that book because of all the help she's given us.

I do want to take a couple of minutes and ask a question that, quite honestly, I know just a little bit about, and I'm wondering if some of the industry members can join in with some of this conversation.

And Kristen, you and I discussed this briefly on the side, and that's the idea of if a company goes through all of the hardship of collecting a reference database, and we've all discussed how many improvements occur with these devices' hardware, one version versus Generation 2, transferring that database into the new system can be extremely difficult for an industry member when they put so much work into that.

If you can just maybe provide a little bit of insight, in addition to what you presented about what the database should look like, what would be a more favorable database that could be transferred to a newer device. I know it depends on that device and how it looks, but just a little bit on that.

And if anybody from the industry groups that are here, if you want to comment on some difficulties with transferring databases from one device to another.

DR. MEIER: What we are looking for, typically, is there could be arguments from an engineering perspective but also a data perspective, to show what kind of differences or bias between models exists. And that would usually include some, at least, small study that would evaluate the old model with the new and looking at what the differences between those models are across the full measuring range of the device.

So, again, some might make arguments on an engineering perspective. But again, I think, for us, it can be hard to make it based solely on an engineering thing, and usually we want to have at least a subset of data.

Again, the standards right here, that you have up, actually the EP9 would be a kind of study that you would do, again, looking, say at 40 subjects across the range. I don't know how prohibitive that is, but that's the ideal, that's what's been developed for lab tests, in transferring reference ranges from one lab device to another.

DR. KAHOOK: Anybody from the audience, from industry, wish any clarification on that? I know it's a point that has been discussed at least in the hallways. It sounds like Kristen might have provided a full answer. Not quite.

Yes, go ahead.

MS. JIL: I'm Jin Jil (ph.) from Optovue.

I've got a question about the transferring database. So when

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

you say about 40 subjects, these 40 subjects must be tested on multiple devices for each of the platforms concerned? Or is it a cross-design or method design?

DR. MEIER: Well, it would be more like a method comparison-type study. It wouldn't be a precision-type study. Obviously you probably already have data on the precision from the cleared device, and then you would need that precision data again with the new model. So I guess -- yes, I'm presuming that you would have that precision data.

And then what I was talking about was more of a method of comparison of, say, comparing means, because you want to really get a good estimate of differences, and to have multiple devices would be great to hone in on what that difference is.

MS. JIL: But these 40 subjects are covering normals and the pathology, when you talk about covering the full range?

DR. MEIER: Probably, to cover the range, you would need that. It depends on the disease, I guess. Again, if we're talking just glaucoma, that's one thing. I know that this technology is used for other things as well.

But yes potentially you would need diseased and normal subjects to really get an adequate coverage of the range. Because, again, if you're going to be transferring, you're looking at normal limits. These are out at the extremes, right? So you really want to be sure you have data out in the extremes, because that's really where you're going to be putting your normal

limits. And if you have a bias in the middle that's very small but out at the ends or extremes of the two models, the two device models are giving very different results, and that's going to have a huge impact on your normal limits.

DR. KAHOOK: So Kristen will be in the room during the break, and I would encourage you to come up to Kristen and ask her the questions.

This is the end of the fourth session. We have about a 16-minute break now to stay on time. So I would encourage you to take a break and then come back to the room at 2:40.

(Off the record.)

(On the record.)

DR. HEUER: Final session on OCT Diagnostic Performance for Glaucoma. I'm proud to be here to with my co-moderator, Bob Weinreb.

And our first speaker is Felipe Medeiros, who will be speaking about "Considerations for Evaluation of Diagnostic Performance."

DR. MEDEIROS: Thank you.

These are my disclosures. They're also in the program.

So I've been asked to talk about implications and how to assess diagnostic performance of imaging instruments.

So what are some of the potential applications of imaging in devices?

So these devices can be used for diagnosis, as we've been

discussing. So when we decrease, we want to decrease the uncertainty in those subjects who are suspected of having a certain condition. So, for example, they're suspected of having damage from glaucoma or suspected of having progression.

It can also be used for screening where we want to identify those cases, they're abnormal or suspected, in the general population or preselected subjects, for example, those in the older population or with certain risk factors such as positive family history or race.

We can also use these instruments for prognosis to determine the risk of developing a condition. I'm not going to talk much about screening because Anne Coleman is going to cover that, so I will talk a little bit about the other two.

So let's start with diagnosis.

And the first important thing about this is that the design of the study should take into account the purpose of the test and should involve the clinically relevant population. And I believe this is something that is frequently ignored actually.

So, again, diagnostic tests are used to decrease the uncertainty about the presence of a condition. So in that, we can use them in a cross-sectional way. So the test is going to be applied at a single point in time in those suspected of having the disease, and I want to see if I decrease the uncertainty that the patient has damage or not. So I'm trying to answer the

question does this patient have glaucoma?

Or longitudinal assessment where the test is going to be applied multiple times during follow-up in suspects or in those with confirmed disease, and I'm trying to answer the question, does this patient have disease progression, which can also help in diagnosis.

So let's start here with the cross-sectional assessment, as the program has been more directed towards this issue here. So uncertainty is what characterizes the glaucoma suspects, so we get a lot of these individuals in clinical practice, and they are those that we are interested in decreasing the uncertainty about diagnosis. So, for example, someone with suspicious optic nerve appearance or an abnormal or suspicious visual field defect. So then we want to use the imaging instrument to help us answer the question does this patient have glaucoma?

So when we look at the literature about imaging devices in glaucoma, we'll see that the vast majority, almost all of them, use a design that consists of cases that are usually glaucoma patients with repeatable visual field defects and controls, which are usually healthy individuals, volunteers from the general population, without any suspicious finding for the disease. That's usually how these studies are designed.

And then the diagnostic accuracy measures are obtained. For example: sensitivity, specificity, or the receiver operating characteristic curves, the ROC curves.

So this is an example of a typical case that is included in most of these studies; for example, a patient with a confirmed visual field defect like this. But why do I need an imaging instrument to help me diagnose glaucoma in this patient? The diagnosis is already very clear. Also, we do not go on applying tests, imaging tests, to diagnose glaucoma in healthy volunteers.

So these case control studies, including just patients with well-defined disease and healthy subjects, they are important for an initial evaluation of the test because obviously, if the test cannot even separate them, it's a poor test and you can just disregard it. But in clinical practice, we use diagnosed tests to actually evaluate patients who are suspected of having the disease, not those with confirmed diagnoses.

So this is a quote from this book, it's a book on evidence-based medicine, which is quite true: "When the diagnosis is obvious to the eye, we don't need any further diagnostic test."

So then these conventional studies, these case control studies with clearly disease versus healthy individuals, the population sample, those included in the studies may not actually be representative of the one that we applied the test in clinical practice. And as a consequence of that, the estimates of sensitivity and specificity or whatever diagnostic measure that you want obtained from these studies may not be directly applicable in clinical practice.

So what we're interested in is whether the test results distinguish patients with and without the disease among those in whom we suspect the disorder. If the sensitivity and specificity actually are obtained in clearly diseased subjects and clearly normal ones, they will be grossly overestimated in most cases.

So let's look specifically at an example here.

So we've been talking a lot about the normative database and how they're constructed. So, for example, if we take the HRT and we see that it gives us the information here, that if you get a red cross, you will have a certain specificity. And how is that done? And we've discussed this before. So you've got a sample of healthy subjects, in general, recruited from the general population, and then let's say if the test is more abnormal, let's say a larger cup area or a smaller rim area, and then you set a limit, let's say a cutoff that corresponds to 1% here so that only 1% of those will be false positives. And that's what the printout gives you, according to the cutoffs, right?

But then the question is should I expect the test to have only 1% false positives in my clinical practice when I apply it to actually the suspect population?

And what's going to happen is the clinical population that we see is going to be enriched by individuals that will be in this area here, those with suspicious findings. So, clearly, the proportion of those falsely abnormal

is going to be much, much higher. So the estimates of specificity that you had before, they do not really apply to that population.

So if the purpose of the test is to complement the clinical evaluation, how should we design the diagnostic accuracy studies?

And I want to do a parallel here to the accuracy of ECG for diagnosing myocardial infarction.

So this is a study by Panju and colleagues. It's a series of papers, publications about how to use clinical tests, called *The Rational Clinical Examination: Is This Patient Having a Myocardial Infarction?*

So they reviewed diagnostic accuracy studies of ECG, and typically, all included patients were suspected of having the MI at the time of the ECG, like it should be. And subsequently the diagnosis was confirmed by another test, let's say the biochemical test.

So the group included here was of patients suspected of MI, not patients, healthy volunteers, that you get on the street or patients who already had the heart attack, no. They were suspected at the point where they were tested with the ECG. Then you have the reference test, and you divide them into those who have the myocardial infarction and those who do not, and then you can measure the accuracy of your test that you're evaluating, the ECG.

But how should we do that? Well, if the target population is of those suspect, then that's the population we should look at.

So we have a patient suspected of having glaucoma, let's say those with high pressure, suspicious optic nerves, but not with clear signs of disease, which would already have made the diagnosis. And then we have the Test X, an imaging test that we want to evaluate, we would have to apply a reference test and divide them in glaucoma and no glaucoma. But I'm sure you're probably saying, well, what reference test? Now we're in trouble because what's the reference that we're going to apply? We don't have the biochemical tests that they have for the MI, so what are we going to do here?

So what reference standard should then be used? Well, based on what I presented, if we use visual fields, that's not a good reference standard because clearly, if they already have repeatable abnormal visual field, they are already diagnosed. I'm not going to apply the diagnostic test to that patient.

So if we want to use cross-sectional optic nerve evaluation, that's also not a good diagnostic test. Why? Because of the huge variability in the optic nerve appearance. Our ability to tell if the patient has glaucoma or not based on a single cross-sectional evaluation of the optic nerve is very poor, so that's also not a very good reference standard.

So then what happens is that these diagnostic studies involving glaucoma suspects will require longitudinal follow-up, either historical or retrospective or prospective, as I will show.

So then let's say we have a cohort of subjects here that are

suspected of having the disease. They have suspicious characteristics that makes us suspect that they may have glaucoma. Then we apply the diagnostic test. One way is applying the diagnostic test here at baseline, following them over time with, let's say, a reference standard, visual fields and photographs, and then dividing them in those who progress to glaucoma and those who do not.

Is this a good design?

First of all, this is a prognostic study. I'm looking to see whether the baseline measurements are prognostic of what is going to happen in the future. But that thing might not yet have happened, actually, here at this point. So this is a prognostic study. So it's important that we assess measures of prognostic ability, which include not only the hazard ratios -- there are not enough -- but some other measures like C-Index, time dependent ROC curves, predictiveness measure.

There are a bunch of different measures that have been proposed for that. And they are very important in quantifying the predictive ability, which is something that the hazard ratios don't do. We need to account for other risk factors, and it's important to keep in mind that a test may have a weak prognostic ability but be a good diagnostic test. And how would that work, for example?

So let's say we have a cohort of ocular hypertensive eyes. They don't have nerve damage, okay? They all have normal nerves. So then it's

not hard to see that if I apply the imaging test here, it's not really going to be able to differentiate them much. They have normal nerves, everything is pretty much normal, they just have ocular hypertension and some other risk factors for glaucoma. But if we follow them over time, some of them will develop glaucoma, nerve damage, and some of them won't.

But, again, because we cannot differentiate them at this baseline here with the imaging, it will have a poor prognostic value for this outcome, right? I cannot really clearly separate these eyes based on this imaging, so I cannot really predict very well what's going to happen because what's going to happen is more influenced by other risk factors, let's say pressure, corneal thickness, age.

However, if I apply the imaging test here, I might actually be able to separate them quite well because at this point, they have the damage, right? So the test might actually have a good diagnostic value if applied here, but will not have a good prognostic value if applied at a certain point in the test.

But how could we assess the diagnostic performance here? Again, remember, we cannot use the cross-sectional evaluation of the optic nerve, and we cannot use the visual fields because that wouldn't be a good reference standard. If they already had the visual field loss, they're already diagnosed.

So we could use the historical follow-up, what happened to

them, so that those who have the progression, we know they have glaucoma. And those who did not have the progression, as long as they were followed untreated and for a relatively long period of time, we know that they most likely do not have glaucoma.

So we know that the process of this progression is actually a highly predictive factor for development of future functional loss. So it's a quite relevant parameter to be used as reference. But then the question that you might say, well, does this all matter really at the end? And that's what I'm going to try to show here.

The influence of the studied population on the diagnostic accuracy. And this is a study that we did with the CSLO. So we had two analyses.

The first one was the conventional one that we tested the ability of the imaging device to discriminate those with visual field loss from healthy subjects. That's the way the vast majority of studies of diagnostic accuracy is done.

Analysis 2, we look to discriminate those suspects who have glaucoma from those suspects who do not have glaucoma. And for this, we had this cohort of suspects, and we differentiated them based on the history of previous optic nerve progression.

And let me illustrate this. So we have a cohort of suspects. So if you look at them here, you cannot really differentiate them. They would

both be suspects that you see in clinical practice. But who has glaucoma here? Can the imaging test help us here? And that's what we see, right, on the daily basis.

So one of them had the test done in 2005 but had a follow-up from 1987, so 18 years without treatment and without any changes to the optic nerve or visual field. So I believe that most of you will be comfortable in saying that this is a suspect, but is likely normal.

On the other hand, this one here was seen in 2002, had a 10-year history of follow-up and had developed progressive nerve changes, progressively lost here in the superior temporal sector, as you can see. So the evidence of progressive damage here confirms the diagnosis.

So then we had these two situations, and we looked at the diagnostic accuracy of several parameters, and I'm going to illustrate one here, but the results were very similar for all of them. I'm going to illustrate with the Glaucoma Probability Score.

In Analysis 1, the conventional analysis, the ROC curve area was 0.89. And if you're familiar with these diagnostic studies done on imaging, you will recognize this as being very similar to what everyone has shown, like close to 0.9, to discriminate normals from glaucoma with field loss.

However, in Analysis 2, we had those who are suspect, and I was trying to differentiate them. Then the performance was actually quite weak. It decreased to 0.65. So that's a huge decrease in the ROC curve area,

as you see these two areas here. So then it does make a difference, it does make a big difference.

And what happens here is that if the patients are referred as suspicious of glaucoma, we get those referrals, by the same characteristic that is measured by the diagnostic test. In this case of the HRT or rim thinning or large cups, these are the patients that we get as referrals as suspicious of disease. Then applying the test, the HRT, will actually bring very little additional value to decrease the diagnostic uncertainty.

For screening, it might be a different situation because then we're testing a population, a general population, where we're trying to identify those abnormal or at high risk, so then it's a different situation. I'm applying the test to a general population, not a population of suspects that have been referred. So then the estimates of diagnostic accuracy, they're obtained from these conventional studies that clearly, glaucoma versus healthy might seem more relevant to these opportunistic screening situations. The diagnostic accuracy estimates are probably more applicable to that situation actually.

And that brings us to the second part here. I'm just going to go through the second talk.

Does it happen with nerve fiber layer, as well?

Can I go on or do you want to break for -- okay.

DR. HEUER: I'll introduce you again.

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

DR. MEDEIROS: Okay.

(Laughter.)

DR. HEUER: Next we have a flash mob doing "Diagnostic Performance Studies": Drs. Medeiros, Chang, and Asrani.

Dr. Medeiros.

DR. MEDEIROS: Okay, thank you.

So let's look, then, what happens with nerve fiber layer.

So if we use the same setup to test the performance of nerve fiber layers, so then you have a cohort of glaucoma suspects followed over time. So you had the retrospective follow-up of what happened to those suspects. You're applying the test today, but you know what has happened to them, but they're suspects today, the date you apply the test.

And they had a follow-up, in this case, of 14 years. Forty-two of them had a history of progressive optic nerve change, so I know that at this date today, if I just look at today's date, there are two suspects because I don't know that history. I need to look at the history to qualify the diagnosis.

So then we had 42 with progressive DON, so they would be the cases, and 86 control eyes, no progression, and followed untreated for this long time period. So we're relatively comfortable in saying that they most likely do not have disease.

So how does the SD-OCT perform, then, in this situation? And then the performance, the ROC curve area for average thickness, was 0.86

compared to 0.72 for rim area.

And probably this happens because in clinical practice, these patients are referred to as suspect because of these characteristics, because of large cups or thin rims. But they're rarely referred based on suspicious nerve fiber layer. So then we don't get that, let's say, biased population. And then the nerve fiber layer parameters in this situation, they actually seem to offer more benefit as a complementary test to the clinical examination, okay?

So we also look at the performance of the SD-OCTs using the conventional analysis: people with glaucomatous visual field defects versus healthy eyes. That's the way most studies are done, again.

And what do you see in the ROC curves of 0.86? So it's interesting to see that they were not that much different than when we use the other design for nerve fiber layer parameters. They were very different, as I showed you before, for the optic disc topography parameters, rim area and cup area, or derivations of that. But for nerve fiber layer thickness, there doesn't seem to be much difference. That's interesting. And I think it's because of the way we actually collect the suspects. Suspects are, in general, those with large cups or thin rims. They are rarely those with suspicious nerve fiber layer.

So the diagnostic accuracies obtained from the conventional studies seem to be similar when we're looking at the nerve fiber layer parameters to those estimates obtained from cohort studies of suspects.

However, disease severity still affects the accuracy of nerve fiber layer parameters, okay? And that can be clearly shown by looking at the relationship between a certain measure of disease severity, which could be, let's say, the degree of visual field loss, going from no field loss to severe field loss and those with no field loss, you can include those with progressive glaucomatous optic nerve damage, so you know they have glaucoma. And then you can apply some modeling techniques, such as ROC regression techniques, to look at the relationship between disease severity and performance of the diagnostic test.

And you can get things like this. For example, this is with the Visual Field Index, which is a measure of -- as you know, a measure of severity of visual field loss. So if you have patients, if you want to differentiate the controls from those with a VFI of 100, so those with no field defect or barely any field defect, the ROC curve was about 0.82. And look at how much larger it gets when you get to more advanced stages of damage. So to differentiate those with more severe damage, it's much easier for the diagnostic test, as it should be expected, obviously. So the ROC curve here gets to 0.96. So you could get, doing this regression modeling technique, you could get your performance for any point of disease severity that you want to detect.

Here is a plot of the sensitivities. And you see how much difference it makes. The dashed one here is the sensitivity for 95% specificity. And you see that for those with no field damage, it starts here at

about .5, so 50%, but then it increases with the increase in disease severity, so the VFI getting lower. And it can get up to 100, over 90 or 100, with those with more severe disease.

But perhaps the most important question is does the SD-OCT nerve fiber layer assessment assist in glaucoma diagnosis? And, again, diagnosis is not about finding the truth, but decreasing the uncertainty that we have.

So let's say when we evaluate a glaucoma suspect, in general, how do we proceed? We proceed with evaluating the clinical evaluation of the optic nerve. Let's say we look at this optic nerve here and then we look at the visual field; the visual field is inconclusive or it's normal; the optic nerve looks suspicious, and I say well, you may have glaucoma, I think maybe you have -- and we do this mentally and subjectively. And we establish, let's say, a pre-test, probability of disease, that in this case I would say it's about 30%.

What we want to know is, after I have done the test, what is the probability? Did it decrease significantly, did it increase significantly? If I have a big change in this probability, it means that the test is helpful. If I do not have a change, it means the test is useless. So I want to bring my post-test probability to very close to 0 or very close to 100, right?

So how can we actually do that, how can we use the result of the imaging test to incorporate in that rationale? So we can use likelihood ratios, and that has been mentioned several times here throughout the day.

So we can get the post-test odds of the disease by multiplying the likelihood ratio times the pre-test odds of the disease, and the odds are just a different way of representing probabilities. So the post-test odds is the pre-test times the likelihood ratio. So then if I have a pre-test as my subjective assessment and I have the likelihood ratio, I get the post-test.

But we've been discussing a lot here about setting some arbitrary limits, and this has been criticized, setting like a 1% cutoff or 5% cutoff, these are all artificial categorizations of the test. But we can actually get the likelihood ratio for a specific test result without relying to those categorizations. So the likelihood ratio is actually the tangent to the ROC curve.

Okay. And this is easy to see, if you remember that one of the definitions of likelihood ratio is sensitivity over one minus specificity, okay? So then there's the tangent at this point, at the limit here, for a particular test result. So then, if you have an ROC curve, you can actually get the likelihood ratio for each point of your test. So then I will know what is the significance of each individual result in the test, what's the significance of getting a 70, a 71, a 72, 73  $\mu$  in the nerve fiber layer. Then I can plug that in that previous assessment I told you.

So, for example, here's a plot of likelihood ratio. This is a work that we have completed recently, calculating the continuous likelihood ratios for each measurement. So as you can see this graph here, you can get from

any nerve fiber layer measurement -- this is for average thickness. You can get the likelihood ratio value. So if you have thin measurement, let's say if you get a 70, the likelihood ratio is going to be very large. So your probability of disease is going to increase a lot if you get that there is no nerve fiber layer.

On the other hand, if you get a nerve fiber layer very thick when you did the test, the likelihood ratio is close to 0, which will bring the probability close to 0. And this is illustrated here, so let's say I have a pre-test here, probability, and I get -- so this continuous line here corresponds to what the post-test probability would be if I had found a nerve fiber layer thickness of 70 in my patient. So from a 0.2, I can increase this post-test probability to actually 0.8.

And you can calculate all of them here in a continuous way depending on the measurement that you get in the test. So going back to this patient. Had 30% pre-test probability. And this involves subjective assessment, which is a limitation we have and we should try, I think we should try, to conduct studies establishing this, how to assess the pre-test probability. And there are a number of publications on that in all the areas of medicine, how to assess the pre-test probabilities.

But then let's say we accept that this is 30. And you could use different numbers and see how it varies. And we do get a test which has a particular nerve fiber layer thickness here, which is 69; I know the likelihood

ratio corresponding to that is 69. And a simple way of actually going from the pre-test probability to the post-test probability is something called the Fagan nomogram. You just go from the pre-test and you cross over here -- and this is an adapted Fagan nomogram -- and based on that -- and you could run these calculations, if you wanted, you would get the post-test probability of disease of 93%. So that suspect had an enormous change in the probability of disease with a particular result here of SD-OCT.

It could have happened that this result of OCT would be a borderline result, okay, and if it was a borderline result, most likely the pre-test probability of disease would not really change much. And that's what happens when we see a borderline result, because the likelihood ratio associated with that is close to 1.

So the diagnostic accuracy studies involving the clinically relevant population can show a clear benefit of nerve fiber layer assessment with spectral domain OCT in decreasing this uncertainty in glaucoma diagnosis. But clearly we need to conduct the studies using the clinically relevant population to show that benefit.

So I have some slides here addressing the questions, but we can leave this for the discussion, if you want, or I can provide my opinions about it if you -- whatever.

DR. HEUER: Why don't we hold them for --

DR. MEDEIROS: Okay.

DR. HEUER: -- discussion and then we'll have them presented during the discussion?

DR. MEDEIROS: Okay.

DR. HEUER: Dr. Chang.

(Applause.)

DR. HEUER: If I could have Dr. Asrani go ahead and come up on the stage so that we can decrease the turnover time, that would be great.

Dr. Chang, take it away.

DR. CHANG: I'd like to start off today by thanking my moderators for the invitation to speak.

We just heard a very impressive talk about how to judge diagnostic performance for retinal nerve fiber layer and the limitations of using our current common study design, such as case control cross-sectional studies. I'm going to be covering the diagnostic performance of the optic nerve head measurements.

I used to do research with Zeiss, one of my financial disclosures.

I'll quickly review how the Stratus algorithm works to calculate rim area. Stratus utilized six radial lines centered over the optic disc, and basically, it was trying to measure a neural retinal rim from each slice, and initially, besides using the ILM and the end of the RPE, an arbitrary offset of 150  $\mu$  was set superiorly to the RPE, and this was considered the inferior

portion of the rim.

But because of eye movement and variability of the scan with imperfect centration, one study revealed that 12% of eyes had imaging artifacts in at least one radial scan due to vitreous opacities, RPE, or small disc problems, and these were in good quality scans. Thus, the optic nerve head was not the favored usage parameter for Stratus.

Now, we move on to spectral domain OCT, and there have been two machines that have released optic disc parameters that have published studies in the literature. And the Cirrus version tried to fix some of the problems of Stratus. So basically now you have a 6 x 6 cube of data which can be summed up and integrated into an en face view, which is basically just a front facing scan. And this allows you to get polar coordinates of the optic disc, as well as look at the ends of the RPE or Bruch's termination relative to ILM, and it also allows you to look at an inner cup defined now as a tangential line to the inner lining ILM. And this is because we think that almost all the RNFL must pass through this border here, so you want to include it at the widest point. This has also been referred to as a minimum distance band concept in Dr. Chen's paper.

So this is part of the method of how Cirrus decides to include the neuroretinal rim. About 180 points are used to determine the rim in a spiral around the entire optic nerve head.

The RTVue machine still uses the Stratus definition of the

neuroretinal rim.

And now we'll just cover the optic nerve head parameters for both machines.

We've discussed how they are repeatable, and the first one is rim area, which is marked in the gray area there.

The next is the disc area, which helps you figure out what size nerve that you have, and typically, we've categorized into large, medium, or small.

Then it also gives you a measure of the cup-to-disc ratio, which is sort of the area of the cup to the area of the nerve, something that we have a harder time doing, but clinically, we sort of assess it in a vertical manner, and that's why we often talk about the vertical cup-to-disc ratio because we can understand linear ratios easier, mentally.

And, finally, there is a cup volume parameter, but this has to use a depth offset as well, and thus is not as commonly used.

For RTVue, they have the same parameters, but the offset is a plane that's defined 150  $\mu$  above the two ends of the RPE.

So when Cirrus was designing the algorithm, they wanted to make sure they took into account where peripapillary atrophy was occurring and how to correct this border.

Now that we've been trying to compare stereo disc photos and experts marking where the edge of the disc is relative to enhanced depth

imaging of the optic nerve with our spectral domain OCT, we've learned that the opening of the neural canal can actually be in different directions, and thus the narrowest portion may be an oblique faction that you can't really see on a stereo disc photo. And thus this is helping to refine our algorithms to make sure that we always find the same point as a determination of Bruch's each time.

Here are some examples of when Cirrus was first starting out to outline the optic nerve head. It was important to make sure that the disc was always centered because if decentration was off, then the polar transformation of the circle around the disc would sometimes not include the correcting.

Is it advancing?

Okay, so here's an example of a tilted disc, and the algorithm was able to figure out that the cup was still small. However, there is a recent study that when looking at high myopes, there was at least 17% algorithm errors due to the severely tilted disc, large peripapillary atrophy that you find in high myopes, as well as a steep slope of the retina surface and vitreous opacities.

Here's an example of a Cirrus OCT from my clinic where the algorithm made a mistake because it initially included the peripapillary atrophy portion and then, by re-centering the disc, you can see the difference there. It happened to have an actually very small oval tilted disc, and with

correction, they would then outline the correct portion.

So given that the optic nerve head algorithm still had some difficulty compared to RNFL, making sure that's always in the right spot, the question for me to answer is how good is a diagnostic performance for optic nerve head parameters?

Well, we have a lot more studies related to retinal nerve fiber layer compared to optic nerve head because these algorithms were only released in Cirrus in Version 5.0 and in RTVue after Version 3.5, so they've been around for just a couple of years. There are four studies with Cirrus that I was able to find and four studies with the Optovue RTVue machine.

Typically, in these studies, we seem to fall along a similar definition of glaucoma using perimetric disease, but some studies try to also include pre-perimetric, but that can be problematic when you're trying to look at a cross-sectional diagnosis.

So here's a summary table of the receiver operating curves or area under the curve values for each of the parameters I listed. You can see that in Mwanza studies, the number was a little bit higher than in the Sung study for rim area and vertical cup-to-disc ratio. This may be due to the type of population that you select because the Mwanza studies happened to use normative database patients, thus may be using super-normals and thus making it appear better than it actually would be in clinical practice.

And this just gives you the same study, sensitivity/specificity at

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

specific cutoffs, since area under the curve usually helps you better when you see the graph.

And here's a table in the same format of early glaucoma or perimetric disease for RTVue, and you can see that the rim area and vertical cup-to-disc ratio numbers were in the .8s, which is lower than in Cirrus, and this may be due to the fact that they use the 150  $\mu$  offset, so it would not be standardizing all types of optic nerves.

And here are the corresponding sensitivity and specificity values, which are a little bit lower. But, again, these values are only applicable for the type of population that you're looking at, so if you do want to just distinguish a true normal from perimetric disease, then the numbers, if you are screening a population like that, would be relevant.

As mentioned before, with the cross-sectional study design, it's difficult to include glaucoma suspects or pre-perimetric disease because typically we need longitudinal data to see who actually converts to glaucoma. Yet, because OCT can theoretically identify disease even earlier than perimetric glaucoma, then we have a problem there because there is no gold standard. It's almost like we need to have some kind of blood test or rate of ganglion cell loss that you could take at one single time point and say that is glaucoma, and then it would make our discussions much easier. Until then, we're going to have many discussions like this about longitudinal studies would be ideal but are much more difficult to carry out.

And then, of course, we can always, as we've shown with severity, it's easier to distinguish normal from glaucoma at more advanced disease, but then what is the worth of the test because then at that point, you don't really need it.

When you're trying to use a sensitivity/specificity cutoff for diagnosis at a single cross-sectional time point, you really would have to have a pretty extensive reference database because of the amount of variability, especially within the optic nerve. So even though RTVue, in their Version 4, tried to expand their database to 861, much larger, there are still more patients than you would be able to figure out on just one test if it's normal or abnormal.

And, finally, this is the last slide.

We're looking forward to the results from longitudinal studies from the diagnostic innovations in glaucoma and advanced imaging for glaucoma, because this will really help us decide what amount of change would be significant. Until that time, basically what I do in clinical practice is I use the spectral domain to take a picture of the time when the patient comes in and that forms their baseline. And due to the high repeatability of the parameters, I look at the most repeatable ones and see if it significantly changes.

Thank you.

(Applause.)

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

DR. HEUER: Next, Dr. Asrani will speak to us about "Macular Measurements."

DR. ASRANI: Thank you, Dr. Weinreb, Dr. Heuer, Dr. Liebmann, for having me here.

Good afternoon, everybody. I apologize for the voice, it's a bit of a sore throat, but I'll try and get through the talk.

I'm the final member of the flash mob this afternoon.

(Laughter.)

DR. ASRANI: We'll be talking about macular thickness in glaucoma.

Here is my financial disclosure.

Why should we be interested in retinal thickness in glaucoma? The ganglion cell layer is multi-layered in the macular region, and along with the nerve fiber layer, constitutes 40% of the total retinal thickness. The majority of the entire retina's ganglion cell population resides in the macula. And, additionally, variation in these macular ganglion cell numbers is small based on the anatomic studies conducted by Christine Curcio.

Unfortunately, as you've already heard, the visual field relatively under-samples the macula. There are about 230 ganglion cells in the center visual field spot compared to only 10 ganglion cells in the periphery. And correspondingly one needs to lose about 70% of the ganglion cells near the center to create a 3 dB loss compared to 44% in the periphery.

Now, macular thickness in normals and in glaucoma was first described by us in the 1990s. We also demonstrated the losses in glaucoma patients. Since then, a variety of macular thickness parameters have been used. Some strategies have included asymmetry of macular thickness, segmented macular ganglion cell complex measurements, increased area of the GCC measured, the ratio of the GCC to the total retinal thickness, as well as hemifield thickness.

The strategy I will describe just now is something that I have used over the past four years. It's called the asymmetry analysis comparing the original thickness measurements between the eyes and between the hemispheres of each eye. This was done because we did not have a normative database of the enlarged area of macular thickness when we started this, and we also found that we could take advantage of the fact that there is tremendous symmetry between the two eyes of an individual. And asymmetry is the hallmark of glaucoma.

So we created an asymmetry display between the eyes, wherefore each small  $3^\circ \times 3^\circ$  area, we subtracted the right to the left and the left to the right and were able to pick out small areas of asymmetry. We also conducted the hemisphere asymmetry test where we subtracted the upper from the lower hemisphere and were able to see differences in 0 to 230  $\mu$  in a particular hemisphere.

You've all heard this morning about red disease. But what

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

about green disease? Why? That's also a big problem. Here is an example, a perfect example, of green disease.

Patient who has a nerve fiber layer defect here, that is demonstrated very clearly in the cross-sectional image, and the dip into the abnormal. But because the quadrant averages the measurement, it gets colored green. This is green disease. These patients will be called normal and they are actually abnormal. But when you look at the corresponding macula, one can make out a very clear-cut loss of ganglion cell layer and macular thickness that extends and becomes actually shallower as it reaches the optic nerve.

I'll share two case examples.

Here is a glaucoma suspect within normal visual field. This is a person who very clearly you can see there is an asymmetry between the two eyes. The right eye is thinner in this region. By just pictorially looking at this, though we are great identifiers of pattern recognition, I would not have been able to detect that there was an asymmetry or a problem between the two, but you do a mathematical subtraction and you can see a loss between the eyes and between the hemispheres. And if you do the 3D reconstruction, you can see the gutter, you can see the groove, where the loss of nerve fiber layer has taken place, right here.

This is to illustrate the importance of ganglion cell losses in the paracentral region versus the periphery. Here are two cups. Where I would

have called this person a glaucoma suspect except that I had the visual fields in front of me that showed me that this person had lost a significant inferior nasal step and this left eye, showing a small paracentral scotoma. Which eye should I be more worried about? What is his outlook? What is his potential? He is only 35 years of age. How many ganglion cells does he have left? Is it going to affect his reading vision? Yes, with macular thickness, I am able to tell that.

There is more loss in the left eye, which has a one small paracentral field effect because the paracentral field loss takes place, only 70% of the ganglion cells are dead. I can tell him that yes, it may affect your central vision because I know that the central ganglion cells are being affected.

And this is to illustrate that in end-stage glaucoma, when all other parameters fail, nerve fiber layer thickness or visual fields, there is still residual macular thickness left over that you can monitor for progression.

In our own study, the inter-visit coefficient of variation was very small for central and perimacular compared to the average RNFL. Even the intra-visit, the macula, was better in reproducibility.

A big thing that has already been talked about today is artifacts affecting the RNFL and macular scans.

Now, macular scans, we all know, are affected by diabetes, macular degeneration, various macular pathologies. But those are relatively

easy to diagnose. The same macular pathologies that affect the macula also affect the peripapillary nerve fiber layer, which we are completely unaware of, to date. The Weiss ring, that happens with the posterior vitreous detachment, that happens to all of us, is exactly overlying the same 3.4 mm circle overlying the peripapillary nerve fiber layer. When the peripheral vitreous, the posterior vitreous detachment takes place until that time it takes place, there is tug on the nerve fiber layer causing it to thicken.

On the contrary, here is myopia with scsis of the nerve fiber layer, which is a very common phenomena. There is a cutoff, truncation, because of the tilt of the optic nerve.

Epiretinal membranes, they don't have to occur only in the macula. They are, in fact, the most common cause of artifacts in RNFL measurements. They result in appearances of thicker-than-normal nerve fiber layer. Macular thickness measurements are also affected, obviously, and this you can recognize by the scalloped nature of this excessive macular thickness when there is significant RNFL loss.

Posterior vitreous boundaries that can cause pretty obvious defects and progression. Here is a patient who shows progression between January 2011 and November of 2011. This is the subtraction. But really or not? If you go and see, you can see that posterior vitreous detachment all the way, the epiretinal membrane resulting in this area becoming thinner. So it was not progression.

Other artifacts are based on other diseases, such as non-arthritic ischemic optic neuropathy that causes complete hemispheric loss, where the loss is greater towards the optic nerve than towards the temporal side.

Hemianopia. Most chiasmal hemianopia up to the lateral geniculate body, strokes are reflected in the ganglion cell layer.

And the Holy Grail, finally, detecting progression. This is real progression. November 2010, April 2011, September 2011. Now we are not waiting for five years. We are waiting six months, twelve months, and we are able to see progression in the nerve fiber layer.

But that's not the purpose of my talk. The purpose of my talk is to show you that the macular thickness also changes. You can see, and we can pick up, easily, arcuate losses on the macular thickness. And when you get two parts of the eye that are correspondingly affected, you can be relatively sure that you're looking at progression, both macular and RNFL.

Here is another example of a similar case. There is progression. And once again, there is an arc-like progression extending from the optic nerve.

This is very subtle progression over a few months. September '09, December '09, and then March '11. Every time the patient came, there was a disc hemorrhage. The nerve fiber layer quadrant average worsened. And if you look at the actual macular thickness cross-sections, you will see the

ganglion cell layer hollowing out. So before it hollows out, it's still structurally intact. Measurements will still be normal. Measurements will become abnormal as time goes on.

What are the findings from recent studies? These are all 2011-2012 studies. Macular ganglion cell parameters were found to be superior to the nerve fiber layer in detecting early glaucoma, especially in high myopes. Both retinal thickness measurements, as well as GCC, are highly reproducible. Both retinal thickness and macular ganglion cell complex show similar sensitivity in progression detection and are better than the RNFL parameters, especially in advanced glaucoma.

I am not here to make the case that the RNFL or the macular thickness, which one is better. We have two of them in the same individual eye. Let's use both together.

So our conclusions are that the high reproducibility of the macular thickness holds promise for objective measurement of glaucoma progression.

Detection of glaucoma and measuring progression is made easier by combining the diagnostic potential of both RNFL and macular thickness.

Artifacts have to be ruled out before accepting results from any technology. As a predictable and significant structural relationship exists between macular thickness and visual field effects, such measurements

should be studied further.

Thank you.

(Applause.)

DR. HEUER: When I first read the next topic, I thought Anne was bringing us a movie, but it's actually a different kind of screening.

And she'll be talking about "'Screening': Can OCT Imaging be Used to Differentiate Normal from Glaucomatous Eyes in the General Population?"

DR. COLEMAN: Thank you.

I'd like to thank the AGS leadership and the FDA for putting on this meeting. I've really enjoyed it today. And I'm going to be really going over screening. I have no financial interest relationships to disclose.

In terms of what is screening, I want to take everybody back to actually the definition. And it's really a strategy that we use to evaluate, in the population, individuals who are unrecognized as having disease and they may have no signs or symptoms of this disease. And so that's something that Felipe Medeiros pointed out, that screening is not diagnosis. We're really trying to have a strategy where we're trying to pick up people who have disease or might have disease and get them then into the clinician's office where they're diagnosed with the disease.

Why do we do screenings? Well, the goal is to treat disease in its early stages if we have a device or a way to pick them up early. In

addition, it's to reduce morbidity or mortality of the disease and to preserve function and quality of life.

There is a well-known criteria for mass screening, and as you can see, in Number 1, the disease should be common and serious, as we feel glaucoma is.

But in addition, for Number 2, the screening test should be highly accurate. And that's what we've been looking at today.

But as you can see, the other three items in the criteria really don't have anything to do with the screening test. It's really that the natural history needs to be understood, there needs to be acceptable treatment and resources for treatment, and also that the treatment should favorably influence outcome.

So a lot of times, when you look at papers or individuals talking about screening, they're not just talking about the test or the tool that's used to do the screening. They're talking about the whole thing, such as using Wilson's criteria.

In terms of when you're looking at the testing, we use sensitivity/specificity, and the one thing that hasn't really been discussed today is positive predictive value, which is the probability that a labeled positive is actually a true case. And it's really the positive predictive value that we like to use a lot when we talk about screening, although it's also important to think about the sensitivity and specificity.

Joshua Stein went over sensitivity and specificity and about the false positive and false negatives. But just to remind you that for more severe disease, we really want to minimize those false negatives, so for more severe diseases, we want to maximize the sensitivity.

And then for those individuals where you have real problems with false positives, because there are economic costs of subsequent exams or there are psychosocial consequences of labeling someone, such as we've seen with CIGTS, with glaucoma, then we really want to maximize the specificity.

In terms of the positive predictive value, just to point out that this is really important because it includes disease prevalence, and that's one of the problems we've had with glaucoma, is that glaucoma is not very prevalent in the general population in the United States. It's ranging around 2%. And so we are talking about screening. With a mass screening, you really need to take some special approaches because you've got a disease that's relatively rare in the general population, although in our offices it seems like it's pretty common.

So when you're looking at screening, sensitivity and specificity are very important for looking at the validity of the test, and it's really the positive predictive value that's going to really let you know is it going to be cost effective to do a screening program with that test.

Now, the Prevent Blindness in America glaucoma advisory

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

board, back in 1998, with Bob Stamper writing for the group, actually came up with some recommendations for what type of criteria would you want for a screening test. And they recommended the test have about 85% sensitivity and 95% specificity, although they felt that 98% was preferable. Now, this was done for screening for moderate to severe glaucoma, so this is glaucoma that has the optic neuropathy plus visual field loss. They didn't really give any advice about how you would screen for very early disease or pre-perimetric disease.

Now, recently, in April 2012, we had the AHRQ Screening for Glaucoma Comparative Effectiveness report come out, and in this report, they said that there's really no one test or a group of tests that would really be suitable for screening for glaucoma at this time. And in this report, they actually looked at a total of 48 studies that we're looking at, the OCT, and really evaluating it for diagnostic accuracy.

So once again, reflecting back on what Felipe Medeiros said, diagnostic accuracy is not screening, and I think that's one of the things that's really important to realize is that with screening, we're not talking always about diagnostic accuracy. And so one of the things that Felipe pointed out is really how do you design studies to do screening.

And when you look at what's come out since the AHRQ report, which finished its data collection about a year ago, October 6, 2011, there's been one study that came out by Boel Bengtsson and Anders Heijl, and they

were really looking at the validity of using the OCT for selective or population screening.

Now, in this study, what they did is they actually had two different populations and they analyzed it separately. But the population we're looking at screening that we're interested in is really the random sample of 307 people that they invited to come to be evaluated because these were 307 people in the general population who were greater than 50 years of age. Unfortunately, they didn't get a great participation rate, only 55%. So usually in epidemiological studies, we like it up around at least 65-70-75%.

They did, however, bring individuals in, and they did do the OCT diagnostic testing, but they also did a complete clinical examination for both individuals who were suspected of having glaucoma who are failing the screening test, but also those who were normal.

One of the things they found is when they used the different cut points, especially the cut point for 5%, is that with both the time domain and the spectral domain OCTs, that the sensitivity for the time domain was about .78 and for the spectral domain was .89. But the specificity for both of them was really very high, about 99% for the time domain and 95% for the spectral domain with the 5% cutoff.

One of the issues with this study, though, is that in the population-based study is that they included in the nine people out of 130

that showed up for the general population evaluation, three of those nine people already knew they had glaucoma. So right there, they shouldn't have been in the screening. You really want just the six people that were newly found out to have glaucoma.

In addition, one of the issues that was pointed out in the AHRQ report, and that you don't see really addressed in this study, is that there wasn't really any masking of the investigators. So when you're doing a screening study, it needs to be done in the general population, you need to mask the investigators to really what the results were with the screening and also then what the results are with the confirmation tests. You don't want the same people doing it, so you want some masking or some barrier to doing that. In addition, you want to make sure that you really describe very clearly who the participants in the study are.

So in this study they have 130 individuals from the general population. We have an idea about the ages. The ones that have glaucoma are in their seventies, about 72. The ones without glaucoma are in their late sixties. But we really don't know how severe the glaucoma is in the six individuals that were diagnosed with glaucoma. So once again, it's really important to really be able to define what's going on in that population when you're doing screening.

In addition, other issues for considering screening with the different devices is not only how good the test is, but also is the device going

to be very easily transportable. You really don't want it falling apart when you put it in a van and then you take it out to start doing screenings that are mass screenings in the general population.

In addition, you want it usable with the standard wall outlet, you want it to be rapid, you want to have few poor-quality scans, you want it to be relatively inexpensive because when you're doing mass screening, you really want to get it out into the community. And so you want the community groups to be able to afford it, but you also want lay people to be able to do it, and you want lay people to really be able to understand what the results mean and then be able to take those patients who fail, who have borderline tests, and refer them to the appropriate clinical offices for evaluation.

Thank you.

(Applause.)

DR. HEUER: While Dr. Rorer is coming up, if I could have the previous speakers and Dr. Jampel go ahead and assemble on the stage so we'll be able to launch right into the panel.

Our last formal talk will be by Dr. Rorer: "Considerations in Characterizing the Diagnostic Performance of a Device."

DR. RORER: Thank you all for coming today.

I'm going to be describing today some general concepts from FDA's draft guidance on "Design Considerations for Pivotal Clinical Investigations for Medical Devices."

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

This draft guidance was issued on August 15th, 2011, and it's available at the link there.

This draft guidance discusses two broad categories of devices: therapeutic and aesthetic devices, as well as diagnostic devices. And obviously, we're going to be concentrating on the general concepts related to diagnostic devices today.

And then I'll be discussing how these general concepts can be applied to imaging devices for ophthalmology but in particular for OCT.

So the first thing that really needs to be considered when trying to design a clinical study that evaluates a diagnostic device is the indications for use. When you're designing a study, it's very important to keep in mind the indications for use when you're selecting the performance measures and choosing the study population. And as been already mentioned, the device should be evaluated in the context of the indications for use.

And what we mean by that is you really need to consider exactly what it is that the device is measuring, how the device is used, when the device is used, for what, on whom, by whom, and what is the device output.

Examples of indications for use, which is our regulatory lingo, you always clear or approve a device for those specific indications for use that are shown here below. These are general categories of indications for use. So these could be imaging only, a qualitative assessment, measurement

of quantitative assessment. This wouldn't mention a specific disease. You could have even the diagnosis of a specific disease, the diagnosis of a specific disease or screening.

And I think it's important to point out here that these last two indications have not been FDA cleared or approved for any of the ophthalmologic OCT devices.

So let's look at these indication categories one by one. Dr. Hilmantel already touched upon this in his talk, but I'm just going to quickly go over some things.

If we're looking at a device for imaging only, only qualitative assessment, what we would like to see is a study that has masked raters that are using pre-established assessment criteria. The images that they would be evaluating would be attained with both a predicate device and the device of interest, the experimental device. The images would have been obtained on the same eyes, in the same location, using equivalent parameters, and there would be numerous pairs across the typical intended population, meaning various pathologies, as well as people free of disease.

The assessment of image quality: An identification of relevant structures and pathologies would be looked at by both of the masked raters and the image evaluated, the image quality evaluated.

The next indication category is measurement or quantitative assessments, and Dr. Hilmantel has already gone over this in his presentation,

so I'm not going to talk about that further.

The next category of indication is aid in the diagnosis of a specific disease. In this type of indication, it has to be very clear that the diagnosis of the specific disease is not made only with the output of the device. It's in conjunction with other information. The sample size of subjects with the specific disease should be large enough to characterize the precision of the measurements obtained in the intended population.

The measurements should not be negligibly affected by a particular disease state, and neither should the variability be significantly affected. So all that has to be evaluated. This may or may not include a normative database, this type of a device that would be seeking such an indication.

For the diagnosis of a specific disease, the device may include a normative database, but it's very important to first establish clinical decision limits, not just normative limits, as Dr. Meier discussed during her talk.

So, first, before you actually conducted the pivotal study, you would have to establish the clinical decision limits. And those limits are how do you discriminate between different states of health, that's what these clinical decision limits allow you to do.

So once you've done that, then you can go on to your pivotal diagnostic performance study comparing the reported diagnosis with that device standing alone without significant amount of other information,

compared to the clinical reference standard. And the study that you're evaluating has to be different than the study that you're doing this evaluation in, though patients in that study have to be different than the population that was used to determine the clinical decision limits.

Again, there have been no ophthalmologic OCTs currently cleared or approved for this type of indication.

Now, for a screening indication, you need to define the intended population for referral and for further workup, because that could be a general population or it could be a somewhat more narrower population. Would it be people just wandering around randomly in a mall, or is it people that are referred to an ophthalmologist's office?

Clinical decision limits need to be clear cutoffs for referral. So that would be how you would define your clinical decision limit for this type of a study. For the pivotal diagnostic performance study for this type of indication, again, you must use a different population than was used to determine the clinical decision limits.

And just to reiterate, no ophthalmologic OCTs have currently been cleared or approved for this type of indication.

So for a diagnostic clinical performance study, again, this is a type of pivotal clinical study for diagnostics, from our perspective at the FDA, where the goal is to characterize diagnostic clinical performance of the device, as well as support its benefit/risk ratio related to this diagnostic

clinical performance.

A diagnostic device is going to be characterized by the clinical performance in such a study, that can be quantified how well the diagnostic device output agrees with the subject's true status as determined by a clinical reference standard. So you need to determine the appropriate clinical reference standard that's been stated over and over during this meeting, and you need to validate the clinical decision limits for the device output in pivotal diagnostic clinical performance study.

So when we think about clinical reference standard, what should we be considering? We should be thinking about the best available methods for establishing a subject's true status with respect to a target condition. This could be a single method or a combination of methods and techniques, and it could include clinical follow-up. You have to keep in mind that clinical reference standards can actually evolve over time, and therefore, it's very important for anyone reporting on the performance of a device and the clinical study of the evaluation of a device to define exactly what the clinical reference standard was that they used for the study.

So what constitutes the best available method? This could be opinion and practice within the medical and regulatory community, there could be several possible methods which could be considered, and again, there may be no consensus reference standard that exists. That's already been pointed out, as well.

So how could you develop clinical decision limits? One way to do it would be to first conduct a cross-sectional study of known normal and known disease subjects of varying severity, and this could provide preliminary information about the possible decision limits. This would allow you to give your distribution measurements, a known normal, and known disease subjects, and hopefully, you'll be able to demonstrate that there is some separation between these two distributions. And then the decision limits that you potentially pick would be between these two distributions.

So once you have your clinical decision limits, an example of a diagnostic clinical performance study that you could do after that would be a prospective longitudinal study. And the goal of this study would be to evaluate the predictive utility of the clinical decision limits. So you would have your clinical reference standard for the disease or condition of interest, and you would enroll representative subjects from the intended use population, for example, glaucoma suspects, and then you would follow the subjects over time.

So the questions that we have and that have already been discussed to some extent during the session are:

5.1. For a cross-sectional study describing the distribution of measurements in known normal and known diseased subjects (excluding suspects):

5.1.a. How should the diseased subject population be defined?

What clinical work-up should be done to establish these populations?

5.1.b. Can the diseased population be further divided by severity of disease using a clinical reference standard and methods? If so, should the clinical performance of the imaging device be characterized separately for each severity group?

5.2. For a longitudinal diagnostic performance study where the structural measurement is taken at baseline and the clinical reference standard consists of a baseline assessment with follow-up:

5.2.a. What subjects should be included in the study population? Should it include glaucoma suspects?

5.2.b. What is an appropriate clinical reference standard?

5.2.c. What minimum follow-up period would provide assurance that the subject has been appropriately classified?

5.3. And with regard to what subjects should be included in the study:

5.3.a. Should there be people other than what we traditionally call glaucoma suspects in that study?

5.3.b. How would you define glaucoma suspect subjects?

5.3.c. What risk factors should we use? Should there be no risk factors?

So that concludes my talk, and I look forward to hearing the discussion of the questions.

(Applause.)

DR. HEUER: We're going to use new slides, which -- we just had the questions elegantly, eloquently, read to us. Elegantly, too, probably.

And we have someone online, but we'll get to that.

So I actually do think that, particularly, Dr. Medeiros' talk framed a lot of these questions for us very, very well. I wonder if the panel wants to weigh in on the question of how, in a cross-sectional study, how should the diseased subject population be defined and what do we use to make those definitions?

Should we put up Felipe's slides? Why don't we go back -- can we go to Dr. Medeiros' slides, the ones he actually queued up some of the answers?

And while we're waiting for that, I think one of the things I struggle with is when we exclude -- and this goes back to some of the earlier questions -- when we exclude people with high refractive errors or funky-looking discs, these are the patients, at least, that I'm challenged with every day, and so trying to apply these tests to those populations where they might, in fact, be the most helpful is often lacking.

I don't know if others want to weigh in on that while we got that or -- here we go. We got it already.

So Felipe, do you want to run us through your answers quickly and then -- you can do it from right there. You can make your microphone

hot and --

DR. MEDEIROS: Okay.

So I believe that it depends -- what are we going to do depends on what the purpose is. So when we do a study that includes a healthy group like the healthy subjects that we've been discussing here, healthy volunteers from the general population versus those with glaucomatous visual field loss, what we're testing is the ability of the instrument to detect patients with glaucomatous visual field loss. And that's what should be very clear.

So, again, as I believe I have shown, it doesn't necessarily mean that when you get those estimates in that particular setup, that they will be applicable to the situation where you're trying to diagnose disease in those who are suspect, that is, those who do not have visual field loss because, again, if they do have visual field loss, then they're not suspects anymore.

So I believe that one needs to start with this type of cross-sectional study because if you show that the imaging instrument doesn't perform well to discriminate these clearly separated groups, then it's because it doesn't really work. But once you get to that step, then you need to give more challenging situations to that diagnostic test, such as the one I described and you have the suspects, a cohort of suspects, and seeing how you can differentiate them.

DR. MEIER: If I could just clarify.

I agree with you that we wouldn't actually view this kind of

data from this study as a diagnostic accuracy study, but I agree, it's a first step. And so, I guess what we were looking at here is what, in this first step, for this kind of study, how would you define what kind of clinical workup would you do to even get the extremes or get that first group of those with disease?

DR. MEDEIROS: If it's cross-sectional, I don't see much other way of doing it unless you do it with a normal population, as we're describing. We can maybe specify a little bit this normal population versus those with a clear diagnosis, which would have to include visual field defects because if you just use a cross-sectional assessment of the optic nerve to diagnose glaucoma, that is certainly not enough, that is certainly not enough.

Would the panel agree with that?

A cross-sectional assessment of the optic nerve, I think it's not enough to define disease by itself, when you don't have field loss.

DR. HEUER: I see Dr. Liebmann itching.

DR. LIEBMANN: No, I think that we all would agree with that, that you need to have field loss, but there are certainly limitations to having those two separate groups, and I think you describe it nicely up there, in your second point.

DR. HEUER: Robert. Dr. Chang.

DR. CHANG: Sometimes I feel it's flawed just to be focusing on field defect because what if the field defect is not progressive? So you can't

really say there is glaucoma unless there's a clinical condition where it's worsening due to pressure-related or whatever other problem. So even saying presence or non-presence of field defect, itself, can be making it difficult to truly know if it's -- you know, there was a glaucomatous event in the past, now it's stable because a secondary effect was removed. So I think to focus on cross-sectional studies for diagnostic purposes is flawed.

DR. HEUER: Dr. Liebmann.

DR. LIEBMANN: I think that leads -- it's a good point, but it leads to, at least, a lot of problems with creating a reference database. I think that when we look at a patient with field loss, you're making an assumption that they started out normal and they developed field loss and they have progressive disease, so they are glaucoma patients. And I think it clearly defines those patients and separates out the ones who have a suspicious disc. So to actually do the reference standard --

DR. CHANG: Right.

DR. LIEBMANN: -- I think you have to make some assumptions. They may not be best assumptions, but they're --

DR. MEDEIROS: Yeah, the other thing I would add is that although these cases certainly do exist -- but I think, when you set up these studies, they will be actually -- if they are in there, they will be such a small proportion that it won't really interfere with the overall assessment, I would think, in this setup, if you have enough sample that you collect that way doing

that.

DR. CHANG: Oh, yeah. I agree with that. It's just, you know, the myopic. You know, we were talking about, like, myopic discs, you need some kind of device to help you, and it's difficult to make a clinical judgment, so those are the situations that I'm referring to.

DR. HEUER: Well, that's actually back to my initial point, and I'll direct this to our FDA colleagues. But one of the things I thought I heard Dr. Rorer saying is that the initial test needs to be set up in the population in which it's intended to be applied, and yet, in many of these studies the population which we end up applying it are specifically excluded.

And I know, with respect to drugs, once it's approved, we can use drugs pretty much as we think is in the best interests of our patients, but what kind of limitation -- did I understand the limitation correctly or the intent, that the approval is based on studying the population in which it is intended to be applied?

DR. RORER: That is correct.

DR. HEUER: Well, I got one right.

DR. RORER: You know, you can come in and say you want this indication, but unless you actually study the device in the population in which it was intended, you're not going to get approval for that indication.

DR. HEUER: Any other comments about Question Number 5.1 and (a)?

(No response.)

DR. HEUER: Actually, (b) goes on to "Can the diseased population be further divided by severity" -- "using a clinical reference method? If so, should the clinical performance of the imaging device be characterized separately for each severity group?"

And I actually thought Felipe used the average --

DR. MEDEIROS: That's next. Yeah, I put it in.

DR. HEUER: Oh, you got it in?

DR. MEDEIROS: Yeah.

DR. HEUER: Okay, good.

DR. MEDEIROS: It's very clear that the performance depends on the severity of the disease, so then it's clear that you need to have a broad spectrum of disease severity and evaluate the diagnostic performance, taking into account that disease severity. And there are several methods that you can do that, and one is using ROC regression methods where you can see very clearly that improvement according to the disease severity. I think that's very clear. The difference is -- certainly, there is a big difference according to the severity, which makes sense.

DR. MEIER: If I could just -- are there defined criteria for those different severity classes? I think that was also one of the things we were seeking from this question, as well.

DR. HEUER: Well, I think that one of the things that we're

always tempted to do is put things in buckets, and what I actually liked about the example that Felipe used, when he was adjusting the presumptive risk based on the findings of average, there was a continuous variable, and I think that's much more powerful, clinically, for us rather than categories, whenever that's possible.

It's obviously not always possible, but -- what would other panelists think?

I know Sanjay might have to leave us early, so why don't we get you to make one comment before we kick you out the door?

DR. ASRANI: I think, certainly, we are stuck currently with perimetry and optic nerve cupping as the two main tests for glaucoma classification, and I think assessment by expert observers of the nerve -- a stereo photograph of the nerve and/or visual field could help us classify patients into different categories. Unfortunately, we have to put them into buckets, like you said, because that's the only way we are currently classifying mild, moderate, and severe glaucoma.

DR. HEUER: Henry.

DR. JAMPEL: Yeah, I think it's essential to differentiate because once a patient has advanced disease, we probably don't need this test, and the issue is how well does this test perform in earlier disease. So unless you separate out early from late, you're not going to have the answer to that.

DR. HEUER: Let's go on to the longitudinal questions. And I

think Felipe may have answers for us up here, too. Yes.

And this is "What subjects should be included in the study population?"

And I think the presumption here is that it should include glaucoma suspects, but the question, if I understand the question correctly, who else should it include?

And go ahead, Felipe.

DR. MEDEIROS: Right, yeah. I gave a simplified answer to that.

I think it depends, the answer depends, on what are you trying to predict there. Are you trying to predict the development of glaucoma, and in the way that there would be a common agreeable definition on that, let's say the development of visual field loss according to certain criteria or the development of progressive nerve damage according to certain criteria, then you can define that as your endpoint.

And you include, obviously, at the baseline, people who do not have glaucoma, who would be a clinically relevant population that you see in practice, which would be the suspects. Of course, it doesn't make sense to do this, including the baseline healthy subjects or those with glaucoma already diagnosed, right?

DR. HEUER: But would it make sense to include some people that you think are clinically normal to see what changes might occur just by time?

DR. MEDEIROS: Well, then that's another issue, right? You're trying to look at -- I think it does, but that's another -- I think it could confound it, also, because how are you going to balance the population?

The way I would design a study like that is a cohort study of people who would be glaucoma suspects, the one that you're going to apply the test, and then you would obtain the measures of predictive ability. That's the way I would -- the issue of how the normals actually change over time, whether they develop changes, that should enter in your assessment of what we call glaucoma, I would think.

DR. HEUER: All right. Go ahead.

DR. MEIER: Again, I guess I'm just -- are there clear definitions of what a glaucoma suspect is?

DR. MEDEIROS: Everyone, right?

(Laughter.)

DR. HEUER: We have about six different definitions up here and about another 100 out there, so --

(Laughter.)

DR. HEUER: So I think they could be --

DR. JAMPEL: So --

DR. HEUER: -- quickly subdivided. Go ahead.

DR. JAMPEL: -- one of the things that I think we need to better define or work out is what we mean by a glaucoma suspect because a lot of

us are a little imprecise, and we've got someone with a family history and they're a suspect, and someone with elevated pressure and they're a suspect, and someone who is older and of African American ancestry and they're suspect, so those are apples, oranges, and something else.

DR. HEUER: Peacocks.

I think either Carlo is exercising an American tradition of the seventh inning stretch or he has a question.

Dr. Traverso.

DR. TRAVERSO: Carlo Traverso. No financial interest in any of the topics.

I'm not sure whether my question really applies to the last two topics, but it seems the most worrisome, clinically speaking, thing is the rate of progression because that's how the patients go into handicap.

How is the panel feeling about really getting that assessed better for what we are doing now, mainly, with visual fields? And it seems, from what I heard today, that the imaging devices are potentially excellent for getting maybe shorter timeframes than with visual field.

The second is that today we all heard a lot of efforts in trying to, still, insisting in identifying pre-perimetric glaucoma patients. And besides the difficulties of pre-perimetry, but is this really that relevant as far as prevention of blindness on a population, identifying pre-perimetric glaucoma rather than looking more efficiently into risk groups and looking at these that

might already have some unquestionable damage in an early stage? I'm just --

DR. HEUER: You know, I think those are really interesting questions. They're a little off-topic for the theme of the meeting, and so we can take that up in the cocktail hour that follows as you run to the airport.

(Laughter.)

DR. HEUER: Gus, did you have a question?

DR. DE MORAES: Yes. Gus de Moraes. No disclosures.

From the prospective prognostic study, which we include suspects, I like the idea that Felipe showed of using change as your, let's say, gold standard to say if that patient has glaucoma or not.

Wouldn't it be the same as looking at the variability at baseline and see if your patient starts closer to the lower boundary of normality versus the upper boundary of normality and you follow these people over time? You got my question?

Because everyone, if you decide to use OCT to follow these patients and this patient is within normal limits but in the lower boundaries of normality, you are most likely going to see progression or convergence of glaucoma in that patient, and that patient will favor your technique that's looking at structure as the best device to detect glaucoma.

DR. MEDEIROS: I'm not sure I'm capturing -- are you? I'm not sure -- you're saying there is a possible -- what you are trying to say is that

there is a possible bias in how you use --

DR. DE MORAES: The question is --

DR. MEDEIROS: -- the reference standard?

DR. DE MORAES: -- we're still discussing normative databases and variability.

DR. MEDEIROS: Uh-huh.

DR. DE MORAES: And we have the upper and lower percentage and it depends on where you are --

DR. MEDEIROS: Right.

DR. DE MORAES: -- on that variability that will actually determine if you're going to --

DR. MEDEIROS: Okay, but then -- so what?

DR. DE MORAES: So your model of looking at optic disc changes over time as your method, as your gold standard, that will correspond to those patients who are actually in the lower boundary of normality when you first looked at --

DR. MEDEIROS: Why?

DR. DE MORAES: Well --

DR. MEDEIROS: Because you're looking -- well -- you need to have, I think, at the baseline a population of suspects that would be -- you don't use to classify -- you're not going to use, as part of your reference, the test that you're using, right, the test that you are evaluating, let's say, the

OCT, will not be part of your reference. That is a very important point.

So your reference, let's say, will be progression based on optic disc photographs or development of visual field loss. So there is no reason to suspect, I believe, that the classification of patients, according to that reference, will be influenced in a biased way by the results of your imaging test. I don't really see how that would happen.

DR. HEUER: We have time for one more question, but it can only be about -- a short question. Sorry.

AUDIENCE: My question is more of a regulatory question. It seems like you can't really define glaucoma, defining a reference is hard, and then finding a diagnosis off of OCT, it gives you a red light or a green light, is not really today's problem, at least not for me.

My question is what is our barrier to bring tools to the research community so that they can share in the sort of analyses that Dr. Medeiros showed and Dr. Asrani showed without having to develop them locally and keep them proprietary and local? So how do we bring the market tools that assist in this research -- is that considered a device in that sense if we do it with that kind of an indication?

DR. MEDEIROS: Let me see. I can answer that specifically to what I presented, like how you could incorporate that in a device to -- for example, I presented a way that you could calculate, for example, the likelihood ratio for each --

AUDIENCE: Right.

DR. HEUER: Right. But this is really a question for the FDA folks.

DR. MEDEIROS: Yeah, okay. All right.

DR. MEIER: Are you asking if you have a new measurement tool, how you could use it?

AUDIENCE: Well, what I'm saying is it is a support tool to support the clinician or the researcher in thinking through all of these diverse elements that go into their diagnosis.

DR. MEIER: I mean, there are research use-only devices, but you're not -- it's not being marketed? I mean, the key is how it's being marketed and sold --

AUDIENCE: We need to sell things.

DR. MEIER: -- and making claims?

AUDIENCE: So my question is we need to sell things. Can you sell a research use-only tool, such as this?

DR. MEIER: No.

AUDIENCE: So how do we support the community answering these complex questions if we cannot --

DR. MEIER: I mean, the OCT devices don't have these kind of indications, and they're all cleared, they're all on the market, so I'm not -- I guess I'm not quite following, unless --

AUDIENCE: Yeah, but the question today is how do you take this to the next step, and so does anyone understand the question?

DR. MEIER: Perhaps Brad Cunningham --

DR. RORER: Brad can try to answer your question. I think I understand.

So the examples of indications that I presented to you, normally, a sponsor will seek those in a stepwise fashion or come somewhere towards the beginning of those indications in my list and then work their way down. So as they develop a better normative database, they develop their clinical decision limits. Then they can come in for those other indications, but -- does that answer your question?

AUDIENCE: It answers. I would just suggest that the FDA might want to consider a pathway to actually commercialize research tools so that we can pay for the development cost and contribute to the community as we're trying to answer questions --

DR. HEUER: Like I said earlier, I think we'll take this conversation into the cocktail hour that follows --

LCDR CUNNINGHAM: I'm sorry. Could I just make one comment to that? This is Brad Cunningham from FDA. There you are.

We do have provisions under --

DR. HEUER: I'm sorry, I'm sorry. We're going to dismiss the panel, and you can keep talking.

LCDR CUNNINGHAM: Okay.

DR. HEUER: And we'll bring Dr. Liebmann and whoever else is -- in closing. So thank you to the panel.

You can go ahead and do the talk now. I'm just trying to get the people off the stage.

(Applause.)

DR. JAMPEL: Malvina, I want to thank you and your entire FDA team for bringing us all together here in Washington. It's been a great experience. We look forward to working with you in the future.

DR. EYDELMAN: Thank you.

While everybody is leaving, let me just make one or two points.

I just wanted to make sure that everybody understands how impressed I am with the outcomes of today's workshop. I truly believe that the input we received from the group today will help us to meet our mission towards glaucoma imaging devices, specifically to facilitate the imaging device innovation by advancing regulatory science and by providing industry with predictable, consistent, transparent, and efficient regulatory pathways.

Today we asked all of you 10 very, very complex questions, and I would like to thank all of the presenters for their thoughtful presentations. I was actually quite impressed by the amount of work that was obviously put in preparation for these presentations.

As well, I would also like to thank all of the attendees for their

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947

thoughtful comments and their input.

As all of you heard, many of our 10 questions got partially answered. Some may require further consensus development in the glaucoma community.

The reason I made these slides is for this purpose. I wanted to make sure that everybody is aware of the link of where today's presentations will be available. So I believe both the slides and everybody's contact information will be available through this link.

Subsequent to that, as Jeff mentioned this morning, we will be working on publication of a manuscript that summarizes today's proceeding.

And we'll look forward to continue the collaborative work with AGS to address the remaining questions.

And last but not least, I once again want to thank all of the people who put this workshop today.

Thank you.

(Applause.)

(Whereupon, at 4:42 p.m., the meeting was adjourned.)

C E R T I F I C A T E

This is to certify that the attached proceedings in the matter of:

FDA/AMERICAN GLAUCOMA SOCIETY WORKSHOP

ON THE VALIDITY, RELIABILITY, AND USABILITY

OF GLAUCOMA IMAGING DEVICES

October 5, 2012

Silver Spring, Maryland

were held as herein appears, and that this is the original transcription thereof  
for the files of the Food and Drug Administration, Center for Devices and  
Radiological Health.

---

CATHY BELKA

Official Reporter

Free State Reporting, Inc.  
1378 Cape Saint Claire Road  
Annapolis, M.D. 21409  
(410) 974-0947